Microenvironment-specific Effects in the Application Credit Scoring Model

Khudnitskaya, Alesia S.

Ruhr Graduate School in Economics, Economics and Social Statistics Institute, Department of Statistics, Universität Dortmund

December 2009

Microenvironment-specific Effects in the Application Credit Scoring Model

Alesia KHUDNITSKAYA

Ruhr Graduate School in Economics, Essen, Germany

Economics and Social Statistics Institute, Department of Statistics, Universität Dortmund[1]

Abstract: Paper introduces the improved version of a credit scoring model which assesses credit worthiness of applicants for a loan. The scorecard has a two-level multilevel structure which nests applicants for a loan within microenvironments. Paper discusses several versions of the multilevel scorecards which includes random-intercept, random-coefficients and group-level variables. The primary benefit of the multilevel scorecard compared to a conventional scoring model is a higher accuracy of the model predictions.

Key words: Credit scoring, Hierarchical clustering, Multilevel model, Random-coefficient, Random-intercept, Monte Carlo Markov chain.

1 Introduction

In retail banking consumer credit scoring plays an important role as a valuable instrument for a decision-making process. Lenders apply scoring models in order to assess credit worthiness of applicants for a loan and forecast the probability of default. This paper contributes to the literature on credit scoring and introduces a new type of a credit scoring model which has a multilevel

[1] khudnitskaya@statistik.tu-dortmund.de

structure. The multilevel scorecard is an improved alternative to a conventional logistic scoring model which is regularly applied in retail banking. In addition, paper proposes a new type of clustering for a hierarchical two-level structure which is more intuitive and efficient in the application to credit scoring. The structure allows exploring the microenvironment-specific effects which are viewed as the unobserved determinants of default. Including microenvironment-effects helps to improve the forecasting accuracy and evaluate the impact of the particular group-level characteristics on the riskiness of borrowers.

In general, multilevel statistical modelling assumes that the data for the analysis is nested within groups. In this set up groups represents the higher-level units and observations are the lower-level units. The structure implies that units within the same group share more similarities than units within different groups. Multilevel models are frequently applied in the field of social science (Steele and Durrant (2009)), political science (Gelman and Hill (2007)) and education (Goldstein and McConell (2007)). In particular, Goldstein (1998) applied a hierarchical structure where pupils are nested within schools to evaluate school effectiveness and compare pupils' achievements between and within schools.

The paper is divided into three parts: theoretical, empirical and discussions. The first section introduces the multilevel structure and explains the motivation for the particular type of a hierarchical structure. In addition, it provides the summary of the data used in the empirical analysis and describes the sources of the data collection. I split the sample into two parts (training and testing samples) in order to compare the out-of-sample performance between the multilevel scoring models and a conventional scorecard.

The empirical part introduces several versions of the multilevel credit scoring models which differ by the composition of random-effects and group-level characteristics. I apply a ROC curve analysis to assess the predictive accuracy of the scorecards and calculate several postestimation diagnostics which check the goodness-of-fit and help to compare the credit scoring models.

Section 4 evaluates the economic significance of the proposed multilevel structure and provides a graphical illustration of the fitted model results. In addition, it shows that the quality of borrowers varies greatly between poor and rich living areas. Applying multilevel modelling allows to account for this heterogeneity.

2 Microenvironment and a multilevel structure

In the multilevel credit scoring the main goal is to define the unobserved characteristics which influence riskiness of a customer additionally to the observed characteristics on borrowers such as income, marital status and a credit history. Accordingly, I define a two-level hierarchical structure for a scoring model which allows including the unobserved determinants of default (random-effects). The structure nests applicants for a loan within microenvironments. In this case the borrowers are the level-one units and microenvironments are the second-level groups. Each microenvironment represents a living area of a borrower with a particular combination of socio-economic and demographic conditions.

There are several reasons why including information on microenvironments in the credit scoring model is important and advantageous. First, it shows that borrowers from dissimilar living areas have exposure to the different risk factors which impact their probabilities of default. It is evident that poor living areas have higher unemployment rates, crime rates, contain a lower share of individuals with a college degree and have a lower level of housing wealth. In such microenvironments individuals have a higher chance to experience an adverse event such as damage of a property, an unexpected income cut or health problems. It is also true that the overall quality of borrowers is lower in low income regions compared to the richer regions which contain fewer borrowers with a derogatory credit history. In this case the microenvironment-specific effects are viewed as the random determinants of riskiness which trigger probability of default. Specifying random-effects and including them in the scoring model improves a credit worthiness assessment of borrowers.

Second, clustering of borrowers within microenvironments allows exploring the impact of the microenvironment-level characteristics on default. In section 3 I evaluate and discuss how area income, real estate wealth and socio-demographic conditions influence the riskiness of individuals within poor and rich living areas.

I define 61 microenvironments within which all borrowers are clustered. The grouping within microenvironments is done according to the similarities in the economic and demographic conditions in the residence areas of borrowers. The economic determinants of grouping are living area income, unemployment rate, purchasing power index and the percentage of department store sales in the total retail sales in the market. The socio-demographic determinants are the share of individuals with a college degree in the living area and share of African-American (Hispanic) residents in the district.

Importantly, the proposed two-level structure where borrowers are nested within microenvironments differ from a standard geographical grouping where individuals are nested in groups according to their geographical locations. The main difference is that the former structure clusters borrowers within microenvironments according to the similarities in the characteristics of their

residence areas. This implies that a particular combination of economic and demographic conditions impacts the riskiness of a customer but not a geographical location itself. Accordingly, within one microenvironment it is possible to have applicants from different areas or cities if their living area conditions are essentially the same.

2.1 Data and variables

In the empirical analysis I use the data from the American Express credit card database which was also applied by W.Greene (1992). The sample contains 13 444 records on the credit histories of the individuals who applied for a loan in the past and for whom the outcome (default or not default) is observed. In addition, I collect the data on the regional economic accounts provided by the Bureau of Economic Analysis (BEA) (www.bea.gov). The BEA data includes annual estimates of personal income, full and part-time employment, taxes and gross domestic product by states and counties.

The individual-level data includes personal information, a credit Bureau report and market descriptive data for the 5-digit area zip-code. I combine the living area descriptive data with the regional-accounts data (BEA) in order to define the microenvironments and create the group-level characteristics.

.

	Full sample	*Training sample*	*Testing sample*
Default	1753	1069	684
Non-default	11691	6997	4694
Observations	13444	8066	5378

Table 1. *Data subsamples*

To compare the out-of-sample performance between the multilevel scorecards and a logistic regression I split the sample into two parts. The short summary of the training and testing subsamples is given in Table 1.

I apply a forward selection approach in order to choose the best performing predictors to include in a scoring model. The resulting set of explanatory variables consists of 12 individual-level variables. Microenvironment-level variables are not included in this set.

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 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 to recognize the microenvironment-specific effects which are defined by the random-effects in the models.

3.1 Microenvironment-specific intercept scorecard

The microenvironment-intercept scorecard extends a logistic scoring model by specifying a varying-intercept at the second-level of the hierarchy. 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 microenvironment. The two-level credit scorecard with a varying-intercept and individual-level explanatory variables is presented in (1). The borrower-level explanatory variables are income ($Income_i$), number of dependents in the family ($Dependents_i$), number of current trade credit accounts ($Trade_{accounts_i}$), a dummy variable for using bank savings and checking accounts ($Bank_i$), number of previous credit enquiries ($Enquiries_i$), an indicator for the high-skilled professionals ($Professional_i$), number of derogatory reports (DR_i), average number of current revolving credits ($R_{credits_i}$) , an indicator variable for the borrowers who have previous experience with a lender such as personal loan or credit card ($Credit_i$), total number of 30-day delinquencies in last 12 months ($Past_{due_i}$) and a dummy variable for the borrowers who own a real estate property (Own_i).

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

$$\begin{aligned} \alpha_{j[i]} &= \gamma_0 + u_{j,0} \quad (2) \\ u_{j,0} | x_{i,k} &\sim N\left(0, \sigma^2_{u_j}\right), \qquad for\ microenvironment\ j = 1..61 \\ Var(u_{j,0}) &= \sigma^2_{u_0} \end{aligned}$$

Given explanatory variables the random-intercept follows a normal distribution with mean γ_0 and variance σ_u^2. The second-level model for the random-intercept includes a population average intercept γ_0 and a second-level residual $u_{j,o}$ as given in (2). The residual $u_{j,o}$ models the unobserved determinants of default which show the impact of the microenvironment-specific effects. The random-intercept accounts for the unobserved heterogeneity in the probabilities of default between borrowers within different microenvironments.

The estimation results for the two-level credit scoring model with microenvironment-specific intercept are presented in Table 2. I fit the scorecard in Stata by applying maximum likelihood.

It is evident, that the coefficient estimates for the fixed-effect variables confirm that the probability decreases with higher income, previous experience with a lender, house ownership and if a customer has both bank checking and savings accounts.

The last row in the table provides the estimate of the standard deviation of the random-intercept. The standard deviation is large suggesting that there is a considerable variation across area-specific intercepts among different microenvironments. On the probability scale the varying-intercept explain changes in the riskiness over and above the population average value by $\pm 15\%$. Importantly, this variability is not explained in the logistic regression scorecard which does not recognize a multilevel structure of the data. Given the normality assumption the 95% confidence interval for the varying-intercept equals $[-2.5; -0.06]$. It shows that 95% of the realizations of the area-specific intercepts are going to lie within this range.

Variable	*Coefficient*	*Std.err.*	*z*	*P>|z|*
Total Income	-0.044	0.004	-9.88	0.000
Number of dependents	0.113	0.032	3.45	0.001
Trade accounts	-0.039	0.007	-5.01	0.000
Bank accounts (ch/ savings)	-0.427	0.082	-5.19	0.000
Enquiries	0.376	0.015	22.48	0.000
Professional	-0.327	0.093	-3.50	0.000
Derogatory Reports	0.622	0.030	20.65	0.000
Revolving credits	0.015	0.004	3.46	0.001
Previous credit	-0.059	0.019	3.16	0.004
Past due	0.239	0.074	3.22	0.001
Own	-0.321	0.109	-2.94	0.006
Constant	-1.270	0.211	-6.01	0.000
Random-effects	*Estimate(Std.err.)*		*95% Confidence interval*	
Standard deviation of intercept, σ_{u_o}	0.61 (0.09)		[0.43; 0.81]	

Table 2. *Estimation results for the two-level credit scoring model with microenvironment-specific intercept. The random-intercept variance and its 95% confidence interval.*

In order to assess the discriminatory power of the multilevel scoring model with a random-intercept I apply a receiver operating characteristics curve (ROC) and calculate several accuracy measures using the curve. In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold. A model with the perfect discrimination has a ROC plot that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC plot is to the upper left corner, the higher the overall accuracy of a model (Zweig & Campbell, 1993).

Figure 1 presents the ROC curve plot for the scorecard 2. Following Hilgers (1991), I also display the 95% confidence bounds for the curve which show the ranges within which the true curve lies. The red triangle on the graph indicates the optimal cut-off point ($c_2 = 0.1376$). This value provides a criterion which yields the highest rate of the correct classifications (minimal false negative plus false positive rates). Importantly, it is possible to define other cut-off points which are optimal according to a specified rule or given a budget constraint. I do not discuss these alternatives in the paper because the decision about an optimal threshold is generally driven by the 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 his budget constraints.

The summary results derived from of the ROC curve and the classification table for the optimal cut-off point are presented in Table 3.

	True		
Classified $c_2 = 0.1376$	D	ND	Total
Default	472	1253	1566
Non-default	212	3441	3812
Total	684	4694	5378
Correctly classified			73.00%
Sensitivity			69.01%
Specificity			73.31%
Area under the ROC (AUC_2)			0.801
Standard error (DeLong)			0.005
95% confidence interval (CI)			[0.794;0.808]
Gini coefficient			0.602
Accuracy ratio			0.688

Table 3. *The summary results for the ROC analysis for the microenvironment-specific intercept model and the classification table for the optimal cut-off point:* $c_1 = 0.1376$

The area under the ROC curve (AUC_2) is 0.8015. This is 0.095 higher than $AUC_{Logit} = 0.707$ for the logistics regression scorecard. The Gini coefficient and the accuracy ratio are also increased ($Gini_{Logit} = 0.368, AR = 0.414$). The 95% confidence interval for the AUC_2 is narrow and does not overlap with the confidence interval for the logistic regression scorecard ($CI_{Logit} = [0.698,0.716]$). The results confirm that specifying a microenvironment-specific intercept improves the discriminatory power of the credit scoring model.

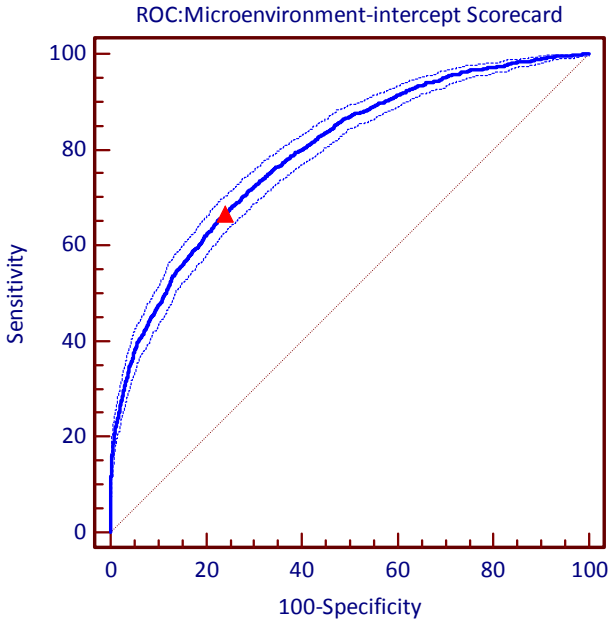


Figure 1. *ROC curve for the two-level credit scoring model with microenvironment-specific intercept. The optimal cut-off point is* $c_1 = 0.1376$.

3.2 Microenvironment-level characteristics in the two-level credit scorecard

In this section I present the extended version of the credit scoring model which allows accounting for the living area characteristics. The credit scorecard is presented in (3). It inserts the microenvironment-level variables in the second-level model for the varying-intercept $\alpha_{j[i]}$. The varying-intercept model is given in (4) includes the population average intercept α_0, the random term $u_{j,0}$ and 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 the microenvironment j. 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 tenth of thousands of dollars, $Stores_j$ -percentage of retail, furniture and auto store sales in the total retail sales in the neighborhood, $College_j$ - percentage of residents with a college degree in the area and $AA_{residents_j}$ - percentage of African-American and Hispanic residents in the region.

Including 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.

$$\begin{aligned} Pr(y_{ij} = 1 | x_{ij}, u_{j,0}) = \quad & Logit^{-1}(\alpha_{j[i]} + \gamma_1 Income_i + \gamma_2 Dependents_i + \gamma_3 Trade_{accounts_i} + \gamma_4 Bank_i \\ & + \gamma_5 Enquiries_i + \gamma_6 Professional_i + \gamma_7 DR_i + \gamma_8 R_{credits_i} \\ & + \gamma_9 Credit_{i_{10}} + \gamma_9 Credit_i + \gamma_{10} Past_{due_i} + \gamma_{11} Own_i) \end{aligned} \quad (3)$$

$$\begin{aligned} \alpha_j &= \alpha_0 + z'\beta + u_{j,0} \\ z'\beta &= \beta_1 Area\_Income_j + \beta_2 AA_{residents_j} + \beta_3 Stores_j + \beta_4 College_j \end{aligned} \quad (4)$$

$$u_{j,o} | x_{i,k}, z_{j,m} \sim N\left(0, \sigma_{u_j}^2\right)$$

$$Var(u_{j,0}) = \sigma_{u_0}^2$$

Similarly, to the previous model (scorecard 2) the area-level residual is assumed to follow a normal distribution with zero mean and variance $\sigma_{u_j}^2$. The two-level credit scoring model with the microenvironment-level variables is fitted in Stata by using maximum likelihood. Table 4 provides the estimation results for the fixed-effect estimates at the borrower and microenvironment levels and the standard deviation for the random-effects.

The estimated coefficients are very similar to the results for the scorecard without the group-level characteristics. This is reasonable as specifying the microenvironment-specific variables modifies only the random-intercept model. The standard deviation of the varying-intercept is decreased which implies that accounting for the second-level characteristics partly explains the variation of the microenvironment-specific effects.

The estimated coefficients for the second-level variables show the impact of the living area information on the riskiness of the applicants for a loan. Higher income in the area has a negative effect on the area-specific intercept. A ten thousands increase in income leads to -0.17 decrease in the intercept . It also true, that microenvironments with a higher share of college graduates predict smaller

probabilities of default. This result is intuitive and shows that the impact of higher education on riskiness 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. Infrastructure of shopping facilities also positively impacts probability of default. One possible interpretation of this result is that good access to the various department stores and shopping malls provokes spending and initiate borrowing.

Variable	*Coefficient*	*Std.err.*	*z*	*P>|z|*
Total Income	-0.041	0.004	-9.34	0.000
Number of dependents	0.114	0.032	3.47	0.001
Trade accounts	-0.038	0.006	-5.02	0.000
Bank accounts (ch/ savings)	-0.426	0.082	-5.19	0.000
Enquiries	0.373	0.015	22.40	0.000
Professional	-0.332	0.095	-3.47	0.000
Derogatory Reports	0.615	0.030	20.51	0.000
Revolving credits	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	0.007
Constant	-0.860	0.21	-4.09	0.000
Microenvironment-level variables				
Living area per capita income	-0.017	0.008	-1.98	0.004
Share of African-American residents	0.012	0.003	3.63	0.000
Share of college graduates	-0.034	0.014	-2.48	0.013
Infrastructure of shopping facilities	0.037	0.029	1.27	0.201
Random-effects	*Estimate (Std.err.)*		*95% Confidence interval*	
Standard deviation of intercept, σ_{u_o}	0.38 (0.08)		[0.24; 0.59]	

Table 4. *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 predictive accuracy of the model is evaluated by applying a ROC curve analysis after the estimation. The ROC curve for the credit scoring model with group-level variables and a varying-intercept is illustrated on Figure 4.4.

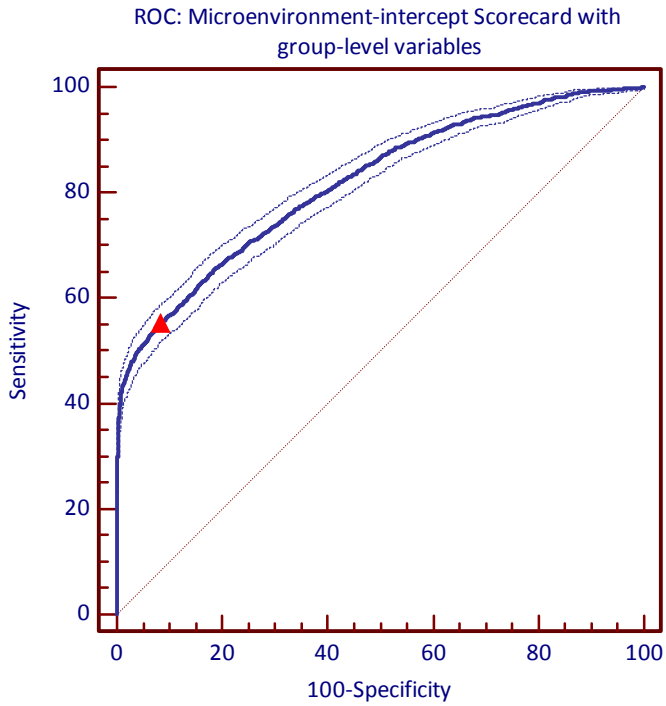


Figure 2. *The ROC curve for the two-level credit scoring model with area-specific intercept and group-level variables. The optimal cut-off point is indicated by the red triangle* ($c_1 = 0.2264$).

The summary results of the ROC curve analysis, Gini coefficient and a classification table for the optimal cut-off point are provided in Table 4.8. The area under the ROC curve and Gini coefficient are increased. The AUC 0.017 is 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]. The standard error of the AUC value is small.

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 (87% versus 75%). This rate is calculated at the threshold which corresponds to the maximal sensitivity / specificity pair ($c_1 = 0.2264$).

	True		
Classified ($c_1 = 0.2264$)	D	ND	Total D
Default	387	384	780
Non-default	297	4300	4598
Total	684	4694	**5378**
Correctly classified			87.21%
Sensitivity			56.1%
Specificity			91.81%
Area under the ROC (AUC)			0.818
Standard error			0.005
95% confidence interval			[0.811; 0.825]
Gini coefficient			0.636
Accuracy ratio			0.726

Table 5. *The summary results 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.*

3.3 Microenvironment-specific coefficients in the credit scoring model

The credit scoring models in the previous sections assign fixed-effect coefficients for the individual-level explanatory variables. This section relaxes the assumption by allowing the coefficients on the two variables to vary across microenvironments. Specifying microenvironment-level coefficients makes a scorecard more flexible and improves the estimation. The area-specific coefficients are viewed as random-effects in a scorecard which show an interaction effect between the borrower and microenvironment-level characteristics.

I specify random-coefficients for the explanatory variable $Enquiries_i$ and $Past\_due_i$. A varying-coefficient of $Enquiries_i$ explains that the impact of credit enquiries on default differ across microenvironments with different economic and demographic conditions. Similarly, the random-slope of the variable $Past\_due_i$ shows that the effect of delinquencies on the credit obligations varies across residence areas of borrowers.

The credit scoring model with the microenvironment-specific coefficients is presented in (5). The second-level models for the random-effects are provided in (6).The model for the area-specific

coefficient β_j^{enq} includes a fixed-effect of credit enquiries (γ_{enq}), a random-term $u_{j,enq}$ and the microenvironment-level variables $z'\beta$. Similarly, model for the varying-slope β_j^{Past} contains an intercept γ_{Past}, group-level variables and a random-term $u_{j,Past}$. The second-level random-effects are assumed to follow a normal with zero mean and variance-covariance matrix Σ_u as shown in (7). Given the individual-level and microenvironment-level variables random-coefficients are allowed to be correlated where ρ is the correlation coefficient.

$$\begin{aligned} Pr(y_i = 1|x_i, z_j, u_{j,Enq}, u_{j,past}) = \; & Logit^{-1}(\alpha_{j[i]} + \gamma_1 Income_i + \gamma_2 Dependents_i + \gamma_3 Trade_{accounts_i} \\ & + \gamma_4 Bank_i + \beta_{j[i]}^{enq} Enquiries_i + \gamma_6 Professional_i + \\ & + \gamma_7 DR_i + \gamma_8 R_{credits_{ij}} + \gamma_9 Credit_i + \beta_{j[i]}^{Past} Past_{due_i} \\ & + \gamma_{11} Own_i) \end{aligned} \quad (5)$$

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

$$\begin{aligned} \beta_j^{enq} &= \gamma_{enq} + z'\beta + u_{j,enq} \\ \beta_j^{Past} &= \gamma_{Past} + z'\beta + u_{j,Past} \end{aligned} \quad (6)$$

$$(u_{j,enq}, u_{j,past}|x_{i,k}, z_{j,m}) \sim N\left(0, \Sigma_u = \begin{bmatrix} \sigma_{enq}^2 & \rho\sigma_{past}\sigma_{enq} \\ \rho\sigma_{enq}\sigma_{past} & \sigma_{past}^2 \end{bmatrix}\right) \quad (7)$$

Table 6 provides the estimation results for the two-level credit scorecard with microenvironment-specific coefficients and group-level variables (Scorecard 4).

The probability of default decreases with higher annual income, number of active trade accounts, if a borrower has previous experience with a lender and if he owns a real estate property. In particular, an average relationship borrower has 1.5% smaller probability than a new customer (no experience with a lender). High-skilled professionals are 8.2% less likely to default. The effect of a house ownership or use of banking deposit accounts is negative. This makes sense as a real estate property or other assets indicate the financial stability of a borrower. These borrowers are 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 garnished by a creditor. Compared to the borrowers who rent accommodation, house owners are 5.1% less risky. Having both checking and saving accounts reduces the probability by 9.53%. At the same time, a derogatory credit history positively impacts the riskiness of an applicant. Additional derogatory remark in the borrower's credit profile increases the probability by 15.1%.

The fixed-effect of the variable $Enquiries_i$ is 0.38 on the logit scale which is similar to the scorecard without a varying-coefficient. On the probability scale the marginal effect of enquiries is 9.5%.

The standard deviation of the microenvironment-specific slope β_j^{enq} is 0.122 which implies that the area-specific slopes differ by $\pm 3\%$ on the probability scale.

Similarly, the fixed-effect coefficient of $Past_{due_i}$ is 0.243. The estimated standard deviation of this coefficient is $\hat{\sigma}_{Past_{due}} = 0.169$ on the logit scale. Translating it to the probability scale shows that the area-specific coefficient explains the change in the probability over and above the population average value by approximately $\pm 4.3\%$.

Variable	*Coefficient*	*Std.err.*	*z*	*P>|z|*
Total Income	-0.037	0.003	-12.43	0.000
Number of dependents	0.131	0.024	5.60	0.000
Trade accounts	-0.037	0.007	-4.96	0.000
Bank accounts (ch/ savings)	-0.384	0.058	-6.56	0.000
Enquiries	0.380	0.021	17.95	0.000
Professional	-0.312	0.100	-3.11	0.002
Derogatory Reports	0.605	0.038	15.81	0.000
Revolving credits	0.011	0.003	2.91	0.004
Previous credit	-0.061	0.017	-3.40	0.001
Past due	0.243	0.053	4.58	0.000
Own	-0.215	0.081	-2.65	0.008
Constant	-1.380	0.100	-13.76	0.000
Microenvironment-level variables				
Living area per capita income	-0.006	0.005	-1.15	0.252
Share of African-American residents	0.008	0.002	3.80	0.000
Share of college graduates	-0.025	0.011	-2.24	0.025
Infrastructure of shopping facilities	0.009	0.007	1.18	0.239
Random-coefficients	*Estimate (Std.err.)*		*95% Confidence interval*	
Std .deviation of β_j^{enq} (Credit enquiries)	0.122(0.019)		[0.089; 0.167]	
Std .deviation of $\beta^{Past}{}_j$ (Past due)	0.169(0.074)		[0.071; 0.401]	
Correlation$(u_{j,enq}, u_{j,Past})$	0.73			

Table 6. *The 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.*

I check the discriminatory power of the credit scoring model with varying-coefficients and group-level variables by plotting a ROC curve as shown on Figure 3. The optimal threshold which yields the maximal true positive and true negative rates is $c_4 = 0.1406$.

The summary results derived from the ROC curve and the classification table for the optimal cut-off point are provided in Table 6. 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 interval for the microenvironment-coefficients scorecard and the interval for the microenvironment-intercept scorecard do not overlap. This confirms that the scorecard 4 outperforms the scorecard 2 and 3 by improving the predictive accuracy. The Gini coefficient and the accuracy ratio are also increased.

I check the discriminatory power of the credit scoring model with varying-coefficients and group-level variables by applying a ROC curve as shown on Figure 4.5. Following Hilgers (1991) I also display 95% confidence bounds for the curve. The threshold which yields the maximal true positive and true negative rates is indicated by the red triangle on the graph.

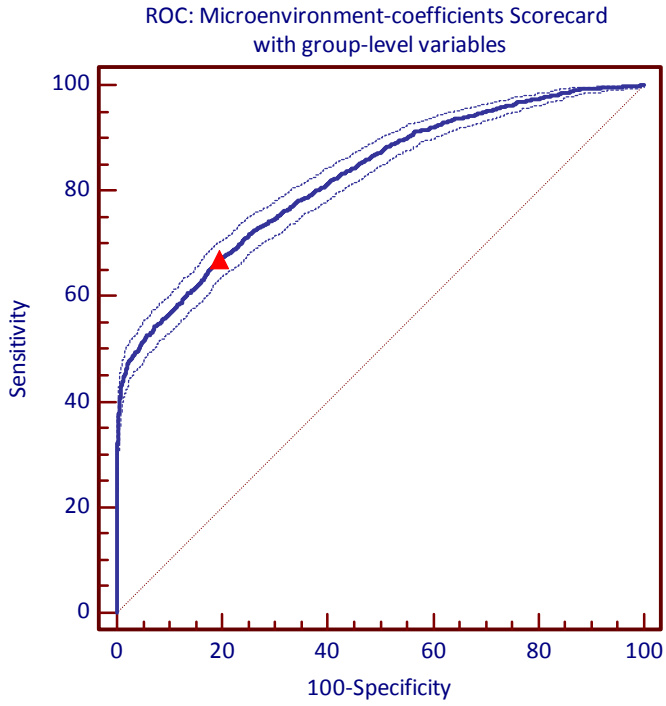


Figure 3. *The ROC curve for the two-level credit scoring model with the area-specific coefficients and microenvironment-level variables. The optimal threshold is* $c_1 = 0.1406$.

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 7. 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 65.8% correspondingly.

	True		
Classified ($c_1 = 0.1406$)	D	ND	Total
Default	450	849	1299
Non-default	234	3845	4079
Total	684	4694	5378
Correctly classified			80.0%
Sensitivity			65.8%
Specificity			81.9%
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			0.741

Table 7. *The summary of the ROC curve analysis results and the classification table for the optimal cut-off point:* $c_1 = 0.1406$.

3.4 Multiple random-coefficients credit scoring model

Section presents a very flexible version of the credit scoring model which includes multiple random-coefficients, microenvironment-level variables and interacted variables. This model extends the varying-coefficients scorecard presented in the previous section. Complementary to the previous structure, I specify two random-coefficients for the individual-level explanatory variables: 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 (8). The microenvironment-specific coefficients are modeled by themselves as shown in (9). The interactions between the borrow-level and microenvironment-level variables are denoted by $k'\delta$ in (8). Interacted variables aim to explain the combined impact of the living area characteristics and individual-level characteristics on the

probability of default. I create three interacted variables which are $Past\_due_i * AA_{residents_{j[i]}}$ - number of the delinquent credit accounts in the past measured at the borrower-level and the living area share of African-American residents measured at microenvironment-level ; $Burden_i * Stores_{j[i]}$ - the access to the various shopping facilities at the area-level and the current credit burden of a borrower; and $Adress_i * Ownership_{Area,j[i]}$ - the share of house owners within a microenvironment and the duration (in months) a borrower stays at his current living address.

$$\begin{aligned} Pr(y_i = 1 | x_i, u_j, z_j) = \; & Logit^{-1}(\alpha_0 + \gamma_1 Income_i + \gamma_2 Dependents_i + \gamma_3 Trade_{accounts_i} \\ & + \beta_j^{Bank} Bank_i + \beta_j^{enq} Enquiries_i + \gamma_6 Professional_i + \beta_j^{DR} DR_i \\ & + \gamma_8 R_{credits_i} + \gamma_9 Credit_i + \gamma_{10} Past_{due_i} + \beta_j^{Own} Own_i + k'\delta \;) \end{aligned} \tag{8}$$

$$\begin{aligned} z'\beta &= \beta_1 Area\_Income_j + \beta_2 AA_{residents_j} + \beta_3 Stores_j + \beta_4 College_j \\ k'\delta &= \delta_1 Past_{due_i} AA_{residents_{j[i]}} + \delta_2 Burden_i Stores_j + \delta_3 Adress_i Ownership_{Area,j[i]} \end{aligned} \tag{9}$$

$$\begin{aligned} \beta_j^{Enq} &= \gamma_{enq} + z'\beta + u_{j,enq} \\ \beta_j^{DR} &= \gamma_{DR} + z'\beta + u_{j,DR} \\ \beta_j^{Bank} &= \gamma_{Bank} + z'\beta + u_{j,bank} \\ \beta_j^{Own} &= \gamma_{own} + z'\beta + u_{j,own} \end{aligned} \tag{10}$$

$$\begin{pmatrix} u_{j,enq} \\ u_{j,DR} \\ u_{j,Bank} \\ u_{j,Own} \end{pmatrix} \Bigg| x_i, z_j \sim MVN(0, \Sigma_u) \quad , \; \Sigma_u = \begin{bmatrix} \sigma_{enq}^2 & 0 & 0 & 0 \\ 0 & \sigma_{DR}^2 & 0 & 0 \\ 0 & 0 & \sigma_{Past}^2 & 0 \\ 0 & 0 & 0 & \sigma_{Own}^2 \end{bmatrix} \tag{11}$$

The random-coefficient model of the variable Own_i in (10) illustrates that the average impact of having a house on the probability of default is $(\gamma_{own} + z'\beta)$. The microenvironment-level residual $u_{j,own}$ 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 second-level residual $u_{j,bank}$, group-level variables and intercept γ_{bank} .

The variance-covariance matrix for the second-level random-effects is constrained to have an independent structure as illustrated in (11). I'm 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. The alternative types of a

variance-covariance matrix specification (such as exchangeable, identity or unstructured) are not discussed in the paper.

The estimation of the credit scoring model in (8) can be problematic with maximum likelihood. The scorecard is complex and contains many of random-effects which should be integrated out in the likelihood. The approximation of the likelihood function can be obtained by applying numerical methods. When the number of the area-specific effects is low a numerical integration produces unbiased estimates. However, the precision decreases as the number of random-effects increases. To solve this computational issue I apply Bayesian Monte Carlo Markov chain (MCMC) to fit the scorecard in (8). This approach is more flexible and more intuitive in the case of a random-effects model where the varying-intercepts and coefficients are viewed as drawn from the population of microenvironment-specific effects.

Variable	*Coefficient*	*Std.err.*	*z*	*P>|z|*
Total Income	-0.031	0.003	-9.92	0.000
Number of dependents	0.133	0.023	5.64	0.000
Trade accounts	-0.031	0.006	-5.16	0.000
Bank accounts (ch/ savings)	-0.368	0.058	-6.28	0.000
Enquiries	0.366	0.013	27.76	0.000
Professional	-0.259	0.098	-2.60	0.009
Derogatory Reports	0.607	0.037	15.85	0.000
Revolving credits	0.005	0.003	2.34	0.020
Previous credit	-0.170	0.068	-2.48	0.013
Past due	0.233	0.050	4.66	0.000
Own	-0.260	0.111	-2.33	0.020
Constant	-1.890	0.280	-6.60	0.000
Microenvironment-level variables				
Living area per capita income	-0.008	0.007	-1.08	0.286
Share of African-American residents	0.011	0.001	5.92	0.000
Share of college graduates	-0.094	0.043	-2.15	0.031
Infrastructure of shopping facilities	0.012	0.005	2.12	0.034
$Past_{due_i} * AA_{residents_{j[i]}}$	0.015	0.019		
$Burden_i * Stores_j$	0.310	0.076		
$Adress_i * Ownership_{Area,j[i]}$	-0.089	0.041		
Random-coefficients	*Estimate (Std.err.)*			*95% Confidence interval*
Std.deviation of β_j^{enq} (Credit enquiries)	0.052 (0.016)			[0.028; 0.100]
Std.deviation of $\beta^{DR}{}_j$ (Derogatory reports)	0.175 (0.085)			[0.068; 0.453]
Std .deviation of $\beta^{Bank}{}_j$ (Banking)	0.048 (0.020)			[0.005; 0.164]
Std .deviation of $\beta^{Own}{}_j$ (Own/rent)	0.664 (0.097)			[0.501; 0.884]

Table 8. *The estimation results for the flexible credit scoring model with multiple random-coefficients, microenvironment-level variables and interacted variables. The standard deviations of the random-coefficients are given together with their 95% confidence intervals.*

The estimation results for the scorecard 5 are provided in Table 8. 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. The large variation is found between the coefficients of the variable Own_i. This implies that the effect of housing wealth considerably varies across areas with different economic and demographic conditions.

The fitted model coefficients of the interactions are not precisely estimated which is not surprising, given I only have 61 level-two groups (microenvironments). The impact of the interaction $Burden_i * Stores_j$ on default is significant and positive. Similarly, the estimated coefficient of $Past_{due_i} * AA_{residents_{j[i]}}$ explains that the impact of the credit delinquencies is higher for borrowers whose living areas contain a higher share of African-American residents. The effect of the interacted variable $Adress_i * Ownership_{Area,j[i]}$ on the riskiness of a borrower is negative. In the richer living areas with a higher level of housing wealth (90% of families own a house) the marginal effect of the length of stay at the address on default is -0.2%.

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 on Figure 4.6. The optimal cutoff-point is indicated by the red triangle on the ROC curve. The 95% confidence interval for the curve is calculated according to Hilger (1991).

The classification table given the optimal threshold $c_1 = 0.1496$, the summary results of the ROC curve analysis, Gini coefficient and the accuracy ratio are presented in Table 9. The area under the ROC curve is increased. It equals 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 testing data sample is not large enough to provide all sufficient 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.

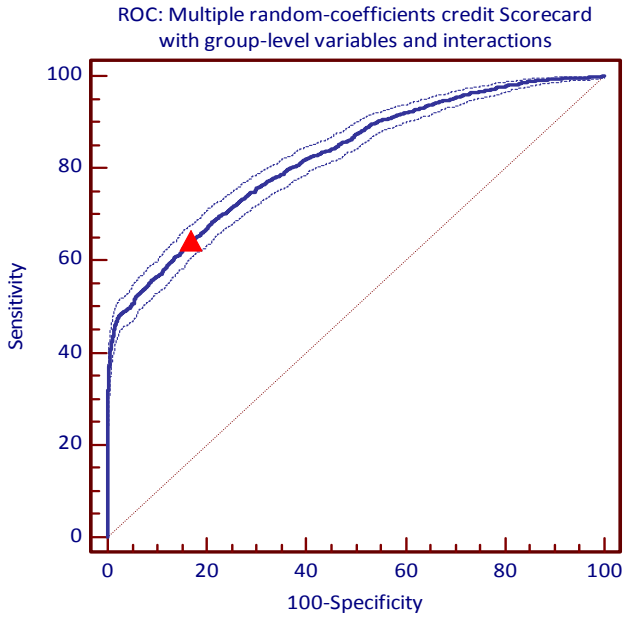


Figure 4. *The ROC curve for the flexible credit scoring model with area-specific coefficients, group-level variables and interactions. The optimal cut-off point is* $c_1 = 0.1496$.

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 weights equally true positive and true negative classifications which may not be the case in retail banking. I discuss the alternative choices for an optimal threshold in the next chapter where I compare a predictive performance between different credit scoring models.

	True		
Classified ($c_1 = 0.1496$)	D	ND	Total
Default	439	778	1217
Non-default	245	3916	3977
Total	684	4694	5378
Correctly classified			81.00%
Sensitivity			64.12%
Specificity			83.42%
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			0.743

Table 9. *The summary of the ROC analysis results, Gini coefficient, accuracy ratio and the classification table for the optimal cut-off point:* $c_1 = 0.1496$.

4 Predicted probabilities and goodness-of-fit check

Section provides several postestimation diagnostic statistics which aim to evaluate the predictive performance of the multilevel credit scoring models.

In general, there are quiet 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, I calculate and report several measures of the goodness-of-fit of an estimated scoring model which are appropriate for a multilevel model and widely applied in the econometric literature. Following Farrell (2004) and Zucchini (2000) I calculate Akaike information criterion (AIC, AICc) in combination with Bayesian information criterion (BIC). AIC and BIC are the tools for a model selection that combine both the measure of fit and complexity. Given two models fitted on the same data, the model with the smaller value of the information criterion is considered to be better. The mathematical details of the calculation of AIC and BIC are provided in Burnham and Anderson (2002), Akaike (1974) and Schwarz (1978).

Section 4.1 summarizes the results derived from the ROC curves for the four multilevel credit scoring models and the logistic regression scorecard. It provides a pairwise comparison of the AUC measures and test the statistical significance of the differences in the AUC values between the credit scorecards. Additionally, I briefly analyse the application of the ROC curve metrics for the evaluation of a scorecard performance in retail banking and describe the alternative measures of the predictive accuracy. 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 4.2 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 score, logarithmic score and spherical score (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.

The graphical illustration of the predicted probabilities concludes the presentation of the fitted model results. It visualizes the microenvironment-specific effects and illustrates the main advantages of the specification of a two-level structure for a scorecard. In addition, I discuss the impact of the microenvironment-level characteristics on default within poor and rich living areas. It is found that economically unstable regions contain a larger share of borrowers with a derogatory credit history and rich living areas have a higher share of borrowers with a good credit history.

5.1 Summary of the 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 on Figure 5.1. The plot combines five ROC curves for the credit scoring models which are presented in chapter 4. The curves are named according to the shortened notation as given in Table 4.12. The logistic regression scorecard is presented by the dashed line and it is assigned the name ROC^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.

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 a 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.

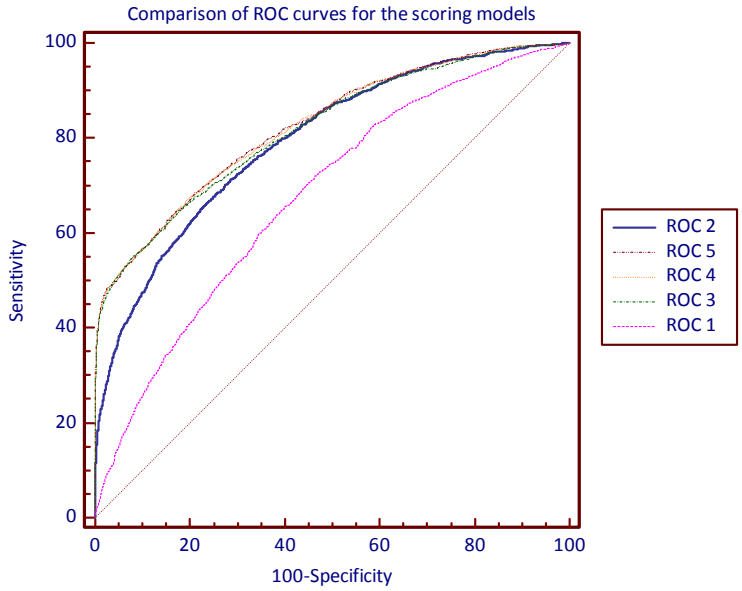


Figure 7. *The comparison of the ROC curves for the different credit scoring models presented in the chapter 4. The ROC[1] for the logistic regression scorecard is illustrated by the dashed curve on the plot.*

In order to give the 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 10. 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 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} is estimated 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 the two AUC values is zero. The test results and the corresponding p-values are presented in the fifths and sixth columns in the table. The 95% confidence interval for the differences in the areas are shown in the forth column in the table.

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

Logistic regression scorecard: area under the ROCLogit curve is AUCLogit=0.707

Table 9. *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).*

The results of the pairwise comparison of the AUC values confirms the statement made earlier that the multilevel scorecards show a higher discriminatory power than the conventional logit model. It is also true that the difference in the AUC measures increases when a scoring model includes more microenvironment-specific effects.

Next, I discuss the relevance of a ROC curve application to retail banking and list the main weaknesses of this approach. In general, a ROC curve is currently considered to be a benchmark approach used to check the predictive quality of a model. It is widely applied in many fields. The predictive performance of a model is measured by computing the area under the curve. However, recently some authors begin to criticize the use of AUC as the standard measure of accuracy (Termansen et al. (2006), Austin (2007), Hosmer and Lemeshow (2000)). They found quiet a few important drawbacks associated with AUC (ROC) measure which prevents its application in practice. In the paper I only briefly discuss the main disadvantages of AUC measure when it is applied in credit scoring and propose the alternative methods.

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. Additionaly, 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 introduce an example when a well-fitted model has a low discrimination power.

A second weakness of the ROC curve and AUC is that they summarise a model performance over the regions of the ROC space 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 are of the ROC curve are valueless.

One solution to the mentioned above weakness would be to compute an area under a portion of the ROC curve. A partial AUC is an alternative to the regular AUC measure which evaluates the discriminatory power of a model over the 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 which yields the highest true positive rate (or false negative rate). The decision about an assessment of a particular region of a ROC curve should be guided by practical considerations within a commercial bank. I illustrate the application of the partial AUC to evaluate a scorecard performance over the region of the ROC curve between two cut-off points.

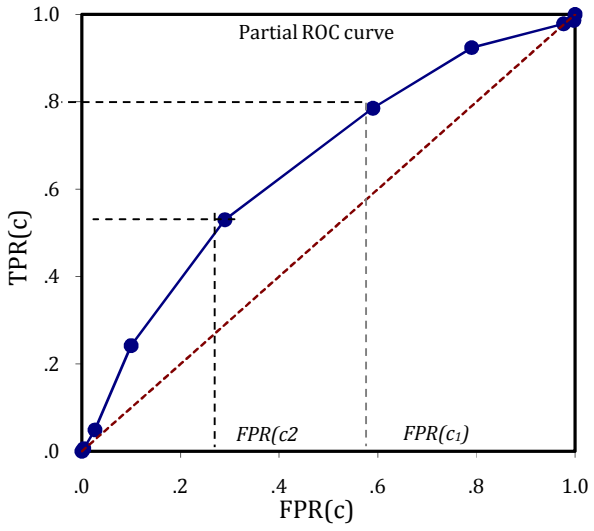


Figure 8. *Partial area under the ROC curve between FPR(c_2) and FPR(c_1).*

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* given a threshold c 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

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

where Y^{ND} and Y^D are the continuos 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 8 provides a graphical illustration of the partial area under the ROC curve between *FPR(c_2)* and *FPR(c_2)* where the c_1 and c_2 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

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

where *FPR(c)* is the false positive rate at the cut-off point c. 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

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

Accordingly, the partial AUC given two cut-off points c_1 and c_2 lies between the maximum and minimum partial areas.

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

The partial areas under the curves are presented in Table 10. 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 a partial AUC ($\frac{pAUC}{pAUC^{max}}$).

Cut-off points :	*[0.1, 0.3]*				*[0.1, 0.2]*			
	$pAUC$	$pAUC^{max}$	$pAUC^{\min}$	$\frac{pAUC}{pAUC^{max}}$	$pAUC$	$pAUC^{max}$	$pAUC^{\min}$	$\frac{pAUC}{pAUC^{max}}$
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 10. *The 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 the relative partial AUC values for the logit scorecard and the multilevel scoring models.*

Results in Table 5.3 confirm that the multilevel scoring models outperform the logistic regression scorecard over the region of the ROC space between two cut-off points c_1 and $c_2(c_3)$. Given the thresholds c_1 and c_2 the scorecard 4 and 5 show similar classification performance. Interesting, given the region of the ROC space between the cut-off point $c_1 = 0.1$ and $c_2 = 0.2$ the scorecard 2 shows the highest predictive accuracy yielding the relative partial area $\frac{pAUC}{pAUC^{max}}$=0.67.

The third important drawback of the AUC value that limits its use as a measure of the predictive accuracy is that it does 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 the 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 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 a true defaulter is $C(D|D)$ and 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 10 times higher than the cost of a misclassified non-defaulter ($C(ND|D)=100,\ C(D|ND)=10$). Table 11 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 point	$c_1 = 0.1$	$c_2 = 0.2$	$c_3 = 0.3$
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 11. *The misclassification costs produced by a credit scoring model given three different cut-off points for the accept/reject decision.*

Concluding the discussion about the application of a ROC curve and derived from it metrics, I suggest that the ROC analysis application to retail banking should be used with caution. In order to evaluate and compare the predictive performance of different scorecards additionaly to the regular ROC curve metrics other measures of accuracy have to be calculated and reported. In particular, the partial area under the curve, misclassification rates and the expected losses given a threshold are the good complements to the regular ROC curve metrics.

5.2 Measures of fit and accuracy scores

In order to compare the goodness-of-fit between the multilevel credit scoring models and the logistic regression scorecard 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, they are applied to select the model which provides the best fit among the range of the fitted models. Table 12 shows the AIC and BIC criteria for the four multilevel credit scoring models and the logistic regression scorecard. The model with the smallest values of both AIC and BIC criteria gives the best fit.

Postestimation statistics	*AIC*	*BIC*
Scorecard 1	2991.34	3090.20
Scorecard 2	2957.18	3062.62
Scorecard 3	2927.17	3045.78
Scorecard 4	2909.24	3041.04
Scorecard 5	2884.50	3029.48

Table 12. *Postestimation 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 by providing a better fit to the data. 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 better classification performance. A flexible version of a scoring model with multiple random-coefficients, microenvironment-level variables and interactions (scorecard 5) provides the best fit.

In addition to the goodness-of-fit check I compute several scalar measures which aim to evaluate the predictive accuracy of the probability forecasts. Following Krämer and Güttler (2008) I calculate Brier score, logarithmic and spherical scores.

The Brier score is the mean squared difference between the predicted probabilities and 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 (13). It shows how large is 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 13.

$$Brier\ Score = \frac{1}{N}\sum_1^N(\theta_i - \widehat{p_i})^2\ ,\quad where\quad \theta_i = \left\{\begin{matrix} 1,\ default \\ 0\ ,\ non-default \end{matrix}\right\} \tag{12}$$

The logarithmic score is another measure of the forecasting accuracy of a model. The calculation of the score is shown in [5.2]. The logarithmic score values are always negative. The scoring rule imposes that a model with the closest to zero logarithmic score shows the best performance. The third column in Table 5.6 presents the values of the logarithmic scores for the credit scoring models.

$$Logarithmic\ score = 1/N\sum_{i=1}^N \ln\left(|\widehat{p_i} + \theta_i - 1|\right) \tag{13}$$

A slightly modified version of the logarithmic score is a spherical score which was introduced by Roby (1965). The calibration of the score is shown in [5.3]. The values of the spherical scores for the credit scoring models are provided in the last column in Table 5.6.

$$Spherical\ score = \frac{1}{N}\sum_{i=1}^N\left(\frac{|\widehat{p_i}+\theta_i-1|}{\sqrt{\widehat{p_i}^2+(1-\widehat{p_i})^2}}\right) \tag{14}$$

Postestimation statistics	*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 13. *The score measures of the 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 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. Similarly, conclusion is made after comparing the logarithmic and spherical scores. According to the spherical scoring rule higher values of the score indicate the model which produces the more accurate forecasts. 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.

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 conventional logit. The goodness-of-fit and the accuracy measures also confirm that the main contribution of the paper is to introduce the multilevel credit scoring model which improves the forecasting quality of a scoring model. In particular, specifying the two-level structure where borrowers are nested within microenvironments and applying the structure to the model results in the efficiency gain. 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 the unobserved determinants of default. Accordingly, including them in the scoring model improves the predictive quality and provides better fit to the data.

Accuracy gain is essential in retail banking where lenders are interested in minimizing the losses associated with lending to bad borrowers (future defaulters).

4.3 Graphical illustration of the fitted model results

4.3.1 Microenvironment-specific coefficients

The credit scoring models introduced in the paper include many microenvironments-specific effects at the second-level of the models hierarchy. The area-specific effects are defined by the random-intercepts and random-coefficients in the scorecards. In order to make the interpretation of the predicted microenvironment-specific effects easier and more transparent I provide a graphical illustration and discuss the variability of the area-specific effects within poor and rich living areas.

Consider the credit scoring model with two random-coefficients which is specified in (5). Figure 9 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 credit enquiries is a very powerful predictor which contains valuable information on the previous applications for a loan. It is assumed that the effect of credit enquiries differ across living areas of borrowers. In the second-level model for the area-varying coefficient $\beta_j^{enq} = \gamma_{enq} + z'\beta + u_{j,enq}$ the residual $u_{Enq,j}$ explains the change in the probability over and above the population average value. The predicted $\hat{u}_{Enq,j}$ are illustrated by the blue points on the plot and the population average effect of enquiries is constant across borrowers and given by the straight red line. Specifying $u_{Enq,j}$ 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 riskiness of a borrower 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 indicate low income areas with a high share of African-American residents and a low level of the per capita real estate wealth.

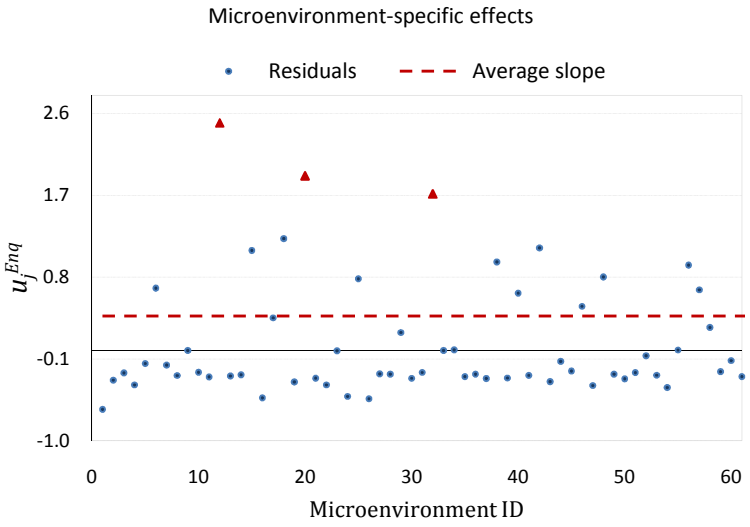


Figure 9. *The second-level residuals of the varying-coefficient of the variable* $Enquiries_i$.*The population average effect of enquiries is illustrated by the straight red line. The abscissa axis is the microenvironment ID.*

If the fixed-effect coefficient is assigned to the variable $Enquiries_i$ then the impact of the one unit change in the number of credit enquires is constant for all borrowers and implies the change in the probability by $\pm 9.25\%$. This assumption may fail 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 kinds of credit products is optimal to offer. Given a residence area retail bankers may choose to 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 the 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 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 very high 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 the situation when the impact of on default is the same for a good and bad borrower which is not realistic in practice. Assigning a varying-coefficient to the variable 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.

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 5.4 illustrates the microenvironment-specific effects ( ) predicted for the five lowest (red charts) and five highest income regions (grey charts). The abscissa axis on the graph shows the predicted measured on the logit scale.

Figure 14. *The microenvironment-specific effects predicted for the five lowest and five highest income living areas.*

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.

4.3.2 Predicted probabilities and living area economic conditions

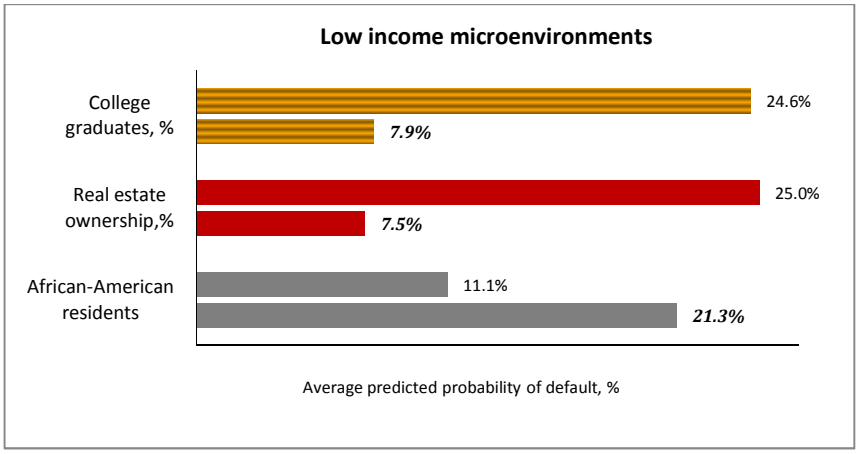
Subsection shows how to apply a graphical illustration of the fitted model predicted probabilities in the postestimation analysis and strategic planning in retail banking. Visualizing the

probabilities not only makes interpretation of the results more transparent, it is also helps to emphasize the role of the microenvironment-level characteristics and explore the impact of the economic and demographic conditions on default.

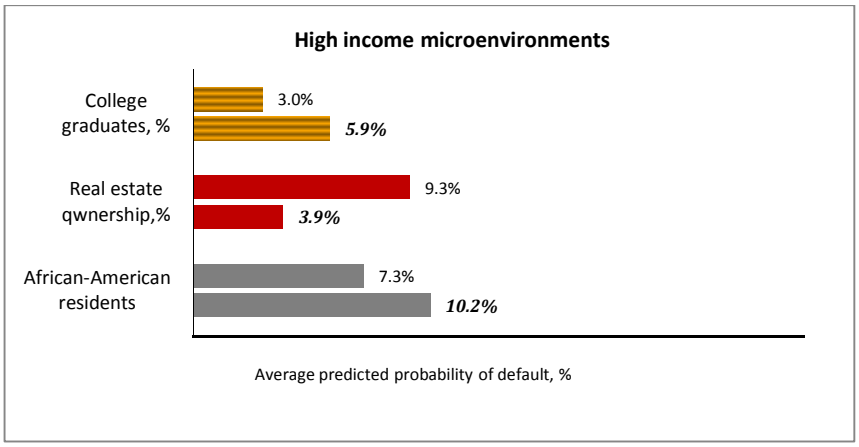
Figure 15 compares the forecasts within the living areas with different economic and socio-demographic conditions. The upper graph *a)* presents the probabilities of default for the low income microenvironment with a high/low share of college graduates in the market (orange bars), with a high/low share of African-American residents (grey bars) and with a high/low 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 financial stabile which implies the average probability of default is 7.5% in these areas.

Controversially, the riskiness increases to 25% if a low income microenvironment also has a low level of real estate wealth (the majority of families rent their accommodation). A high presence of college graduates on 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% which is 16.7% smaller than the similar result for the 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 the high income areas with a high share of African-American. It 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 the area is low .



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



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

Figure 4.11. *Average predicted probabilities for microenvironments with different economic and socio-demographic conditions.*

In summary, the graphical illustration of the predicted probabilities not only shows the impact of the economic and demographic conditions on default, it also reveals that exposure to risk

within high and low income areas also depends on the other living area characteristics such as the real estate wealth, share of African-American residents and share of college graduates. Therefore, clustering of borrowers within microenvironments in the credit scoring model allows to define the effect of the particular combination of living area conditions on default.

Applying a graphical illustration of the predicted probabilities is very advantegeous for a strategic planning in retail banking. It helps to detect the areas where the exposure to the unobserved determineants of defaut is high. Given this information a lender can adjust his market strategy.

Conclusion

Paper discusses several versions of the multilevel credit scoring models which has a two-level hierarchical structure. The hierarchical structure of the model nests borrowers within microenvironments according to the similarities in the economic and demographic conditions in their living areas. The microenvironment-specific determinants of default are explained by the random-effects in the scorecards which are included at the second-level of the hierarchy. Specifying random-effects improves the classification performance of a scorecard and explains the variability in the probabilities of default between the living areas with dissimilar conditions. Additionally, the two-level structure allows exploring the impact of the group-level characteristics such as area income or unemployment on the riskiness of borrowers. Given the ROC analysis results and goodness-of-fit tests it is evident that the multilevel scoring models outperform a logistic regression scorecard. They provide higher predictive accuracy and better fit the data.

The graphical illustration of the fitted model results confirms that within low income areas the probabilities of defaults are higher in microenvironments with a low share of college graduates, low level of housing wealth and high share of African-American residents. The opposite conclusions are made with respect to the high income areas.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* , 19 (6), pp. 716–723.

Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation.* Oxford UniPress.

Baker, S. and, Pinsky P. (2001). A proposed design and analysis for comparing digital and analog mammography: special receiver-operating characteristic methods for cancer screening. *Journal of the American Statistical Association*, 96, 421-428.

Burgess, S., McConell, B. and Goldstein, H. (2007). Modeling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society,* Series A, Vol.4, pp. 941-954.

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

Draper, D. (1995). Inference and hierarchical modelling in the social sciences. *Journal of Educational and Behavioural Statistics,* 20(2), 115-147

Durrant, G. and Steele, F.(2009). Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys. *Journal of the Royal Statistical Society, Series A*, 172, pp. 361-381.

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

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

Goldstein, H. & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society*, Series A, 159:505 13.

Greene W. (1992). A statistical model for credit scoring. *Working paper*.

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

Hilgers, R. A. (1991). Distribution-free confidence bounds for roc curves. *Methods of Information in Medicine*, 30, 96–101.

Jang, M., Lawson, A., Browne, W. & Lee, Y. (2007). A comparison of the Hierarchical likelihood and Bayesian approaches to spatial-temporal modeling. *Environmetrics* 18, 809-821.

Kreft, I and de Leew, J. (1995). The effects of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, 30, 1-21.

McConell, B. Burgess, S and Goldstein, H. (2007). Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society*, Series A, 170, 4, 941-954.

Pepe, M. and Cai, T. (2003) Semi-parametric ROC analysis to evaluate biomarkers for disease. *Journal of the American statistical Association*, 97, 1099-1107.

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

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

Rabe-Hesketh,S. & Skrondal, A. (2004).Generalized multilevel structural equation modeling. *Psychometrika*, 69, 167-190.

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W. & Lunn, D. (1994, 2003). BUGS: Bayesian inference using Gibbs sampling. MRC Biostatistics Unit, Cambridge, England. www.mrc-bsu.cam.ac.uk/bugs/

Steele, F., Goldstein, H. & Browne, W. (2004). A general multilevel multistate competing risks model for event history data, with an application to a study of contraceptive use dynamics. *Statistical Modelling*, 4, 145-159.

Steele, F. and Goldstein, H. (2006). A multilevel factor model for mixed binary and ordinal indicators of womens status. *Sociological methods and research,* 35, 137-153.

Teitler, J. & Weiss, C. (2000). Effects of neighborhood and school environments on transitions to first sexual intercourse. *Sociology of Education*, 73, 2, 112-32.