

# Adverse Selection in Application Credit Scoring through the Prism of Hierarchical Multilevel Modeling

Alesia Khudnitskaya

Ruhr Graduate School in Economics  
Economics and Social Statistics Institute, Department of Statistics, Universität Dortmund<sup>1</sup>

## Abstract.

This paper discusses implementation of a multilevel statistical model in the process of an application credit scoring. The analysis is done in two steps. First, the hierarchical clustering is obtained and applicants' credit histories are assigned to clusters. Thereafter a multilevel random-effects model is fitted.

The model hierarchy treats applicants for a credit as level-1 units which are nested within micro-socio-environments, level-2 subjects. It is supposed that a living environment of applicants plays a substantial role in determining exposure to different risk factors and triggering default cases. Taking into account this unobserved heterogeneity across clusters (microenvironments) leads to more accurate assessment of probability of default and makes possible to estimate cluster-specific effects. The credit scoring model is specified to have a two-level nested structure.

The results confirm that additionally to the customer's financial stability default on a loan depends on economic and socio-demographic characteristics of his living environment. It is shown that applicants belonging to the same high income category but identified by different microenvironments have different exposure to risk mainly because surroundings matter.

**Key words:** credit scoring, hierarchical clustering, multilevel model, random-coefficient, random-intercept.

**JEL-classification:** G21, C53, D14

## INTRODUCTION

In the retail banking and Microfinance consumer credit scoring plays a substantial role as a valuable instrument for a decision-making process. Credit scores help to answer the question how risky is a borrower and make a probability of default forecast.

The main idea of this paper is to illustrate an implementation of a multilevel statistical modeling in the process of application credit scoring. The primary benefit of improved efficiency of a scorecard is higher accuracy of the model prediction.

The analysis is done in two steps. First, hierarchical clustering is obtained and all applicants are assigned into eighty one groups that represent micro-socio-environments they live in and then clustered data is used to create the two-level structure. Model hierarchy treats applicants for a credit as level-1 units which are nested within micro-socio-environments, level-2 subjects.

The concept of two-level formulation is that applicants represented by dissimilar micro-socio-environments face different triggering default cases and have exposure to environment-specific risk factors. Accordingly, it is reasonable to account for this grouping while assessing or forecasting

---

<sup>1</sup> khudnitskaya@statistik.tu-dortmund.de

applicant's probability of default. It is possible to do by specifying a multilevel nested structure that allows bringing unobserved heterogeneity across microenvironments in the model.

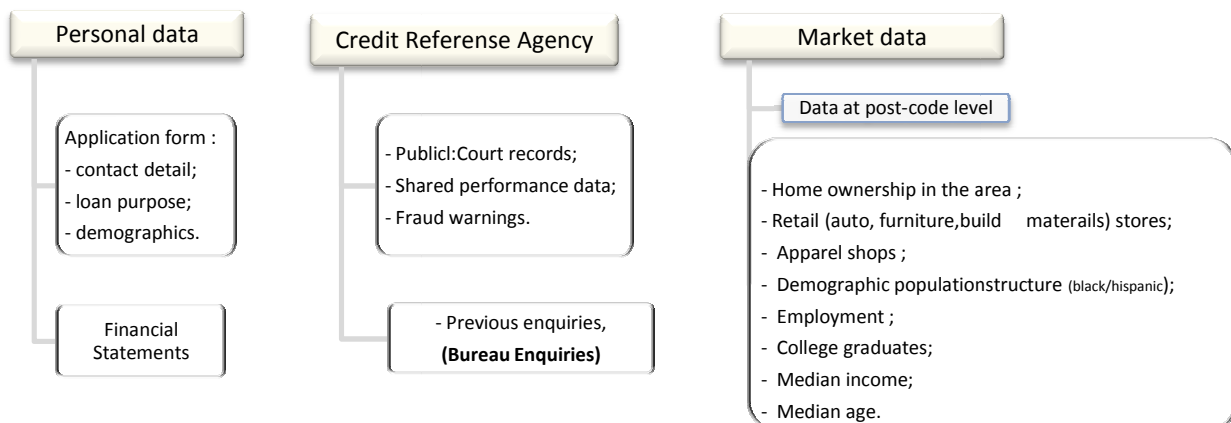
Multilevel models with regression structure are quite extensively used in the statistical literature and very frequently applied to solve problems in social science, medicine and epidemiology<sup>2</sup>. We focus on a random-effects hierarchical model for discrete data which is a powerful tool for accessing between-group heterogeneity.

Generally, for the assessment process financial intermediaries apply particular ratings algorithms in order to rank customers or divide them into groups like "Good" or "Bad". In consumer credit lending it is Scorecards or predictive scoring models. Among the most widely used techniques for scoring are linear and quadratic discriminant analysis (LDA, QDA), generalized linear regressions. This is a class of parametric models that requires certain assumptions hold true. The other set is non-parametric techniques: neural networks, genetic algorithms and K-nearest neighborhood.

## II. Data description and variables selection

The dataset includes 10420 observations and it is a part of American Express Credit applications database. It is a subsample of customers' Credit History data used by W.Greene (1992). The explanatory variables vector consists of 52 covariates and contains applicant's personal data, Credit Reference agency report and market descriptive data for the 5-digit area zip code in which applicant lives (see Diagram 1). Sample estimate of probability of default is 9.5%.

Diagram 1. Types of data



Personal data is collected throughout the fill-in forms and other auxiliary references. It includes such items as age, education, income, additional income, occupation field, number of dependents, home ownership, employment duration, etc. Credit Bureau information collects past credit history of

<sup>2</sup> For more details: 'Multilevel modeling for binary data' Guang Guo, Hongxin Zhao (2000)

individuals and provides data on derogatory reports, number of credit file searches or enquiries<sup>3</sup> and court records.

Descriptive market data is used to perform a hierarchical clustering. Applicants for a loan are nested within clusters according to the similarities in their environment characteristics. From available descriptive data we choose percentage of home ownership in the area, median income, percentage of black/hispanic population in the district, percentage of college graduates, employment and retail infrastructure index be determinants in the cluster analysis. After clustering is done all individuals are assigned to eighty one groups<sup>4</sup>.

We also rescale and normalize some explanatory variables and create some new ones. ( Table 1)

**Table 1**

Created explanatory variables

Created Variable	Description
Infrastructure	– weighted index that show degree of development of certain area, include percentage of retail, apparel, furniture, building materials stores and buy power index.
Bank advance	– represents individual experience in using different banking and financial products: number of checking (saving accounts , other debit cards), department store or gasoline credit cards.
Past due	– describes total number of trade lines 30 days past due and number of 30 day delinquencies in during the last 12 months.
Trade accounts	– number of open and current trade accounts

### III. Dichotomous response multilevel model

#### Single-level model

In this research we deal with the dichotomous response data and apply a generalized linear probability model to build a multilevel structure. From the credit history data we observe occurrence of default if response variable equal 1: customer failed to complete required payments within a six months interval, and 0 otherwise.

Single-level model for logistic regression uses pooled data and assumes that the expected value of the dependent variable equal probability of customer's default given vector of covariates – exogenous predictors  $x$ .

$$E(y_i | x) = Pr (y_i = 1 | x), \quad x = (x_1 \quad \dots \quad x_K)' \text{ - set of explanatory variables}$$

In case of dichotomous response models we require probability of default lie in the interval  $\{0, 1\}$  therefore the link-transformation function  $f(w)$  should be specified:

$$Pr (y_i = 1 | x) = f(\beta_0 + \sum_{k=1}^K \beta_k * x_{ik})$$

<sup>3</sup> Consumer Credit Enquiries- a notice on credit report that details ones attempt(s) to apply for new credit ( mortgage, auto loan, or credit card). Credit inquiries show up on your credit report (whether approved or not) so other creditors can determine if you've been trying to secure new lines of credit recently, which research determines can lead to greater credit risk.

<sup>4</sup> As soon for this paper number of level is set to two but it could be extend to higher number of hierarchical levels or even cross-classified levels

$$(f^{-1}(\beta_o + \sum_{k=1}^K \beta_k * x_{ik})) = \beta_o + \sum_{k=1}^K \beta_k * x_{ik} = \omega_i^{Linear},$$

*where  $\omega_i^{Linear}$  – linear part of the model*

The distribution of the response is independent given explanatory variables and assumed to be Bernoulli ( $\pi_i$ ) or Binomial ( $1, \pi_i$ ). We use logit model as link function:

$$\Pr(y_i = 1 | x) = \text{logit}^{-1}\left(\beta_o + \sum_{k=1}^K \beta_k * x_{ik}\right) = \frac{\exp(\beta_o + \sum_{k=1}^K \beta_k * x_{ik})}{1 + \exp(\beta_o + \sum_{k=1}^K \beta_k * x_{ik})} \quad (1)$$

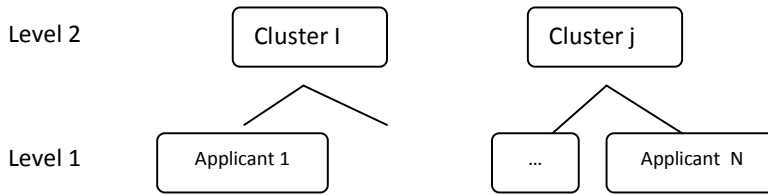
The formula (1) gives the inverse logit prediction of the probability of default given the set of explanatory covariates  $x_k, k = \overline{0..K}$ .

### Multilevel model

We now turn to the multilevel structure of the scoring model. This type of statistical model illustrates extension of a general logistic regression and has a hierarchical structure as lower level units are nested within higher level clusters  $j = \overline{1..J}$  that represent microenvironments people live in (see diagram 2). Socio-demographic and economic conditions of customers' living environments are different and play a substantial role in triggering a default event.

The information available for individuals is represented by a vector of level-1 explanatory variables  $x_{ij} = \overline{x_{1,ij} \dots x_{K,ij}}$  for applicant  $i$  within cluster  $j$  (total income, age, number of dependants, etc). Group level covariates  $z_j = \overline{z_1 \dots z_M}$  represent common for all individual in cluster  $j$  information. It includes market and socio-demographic data (area income, median age, unemployment rate).

Diagram 2. Two-level model structure



For more transparent presentation the two-level mixed-effects model is specified using a two-stage formulation. The varying-intercept and varying-slope model for the probability of default with one explanatory variable  $x_{ij}$  (for example, income of the person<sub>*i*</sub>) is written:

$$\Pr(y_{ij} = 1) = \text{Logit}^{-1}(\alpha_j + \beta_j x_{ij}) \text{ for individuals } i = \overline{1..N} \quad (2)$$

The group-level random intercept  $\alpha_j$  and random slope for income  $\beta_j$  are themselves modeled as :

$$\alpha_j = \gamma_0 + u_j, \text{ for cluster } j = \overline{1..J} \quad (3)$$

$$\beta_j = \gamma_1 + u_{1,j}, \text{ for cluster } j = \overline{1..J} \quad (4)$$

with error terms  $u_j, u_{1,j} \sim MVN(0, \Sigma_{u_j})$  that follow bivariate normal distribution with mean zero and variances  $\sigma_{u_0}^2, \sigma_{u_1}^2$ .

The crucial point here is that we cannot estimate equations (3) and (4) on their own because the random effects  $\alpha_j, \beta_j$  are not observed. Therefore, these level-2 models should be included into level-1 model to obtain reduced form model for observed response  $y_{ij}$ .

Generally, for the set of  $K$  explanatory variables  $x = (x_1 \dots x_K)$  at individual level and  $M$  explanatory variables  $z = (z_1 \dots z_M)$  at group level the reduced form structure with random intercept and one random slope  $\beta_s = \gamma_s + u_{j,s}$  for explanatory variable  $x_s$  is following:

$$\begin{aligned} \text{logit} \left( \text{Pr} \left[ y_{ij} = 1 | x_{ij}, z_j, u_{j,0}, u_{j,s} \right] \right) &= \quad (4) \\ &= \underbrace{\gamma_0 + \sum_{k=1}^K \gamma_k x_{ij,k}}_{\text{Fixed}} + \underbrace{\sum_{m=1}^M \theta_m z_{j,m} + u_{0j} + \sum_{s=1}^K u_{j,s} x_{ij,s}}_{\text{Random}} \end{aligned}$$

Fixed part:

$K$  – number of covariates at a lower level,  $x_k = \overline{x_1 \dots x_K}$  for applicant  $i = \overline{1..N}$  and cluster  $j = \overline{1..J}$

$M$  – number of covariates at a higher level – 2,  $z_m = \overline{z_1, \dots, z_M}$ ,

Random part:

$u_{0j}$  – random intercept at a higher level – 2,

$u_{j,s}$  – random coefficients at a higher level,  $1 < s < K$

It is assumed that given  $\pi_{ij} \equiv \text{Pr}(y_{ij} | x_{ij}, z_j, u_{j,0}, u_{j,s})$ , responses  $y_{ij}$  are independently distributed as:

$$y_{ij} | \pi_{ij} \sim \text{binomial}(\text{Den}_{ij}, \pi_{ij})$$

For the two-level model second level random-effects follow a multivariate Normal distribution  $u_{j,0}^{(2)}, u_{j,s}^{(2)} | x_{ij,s} \sim MVN(0, \Sigma^{(2)})$  with zero means and variance-covariance matrix  $\Sigma^{(2)}$ :

$$\Sigma^{(2)} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{s,1} \\ \sigma_{s,1} & \sigma_{s,s}^2 \end{bmatrix}, \quad \sigma_u^2 = \text{Var}(u_j)$$

This is the two-level model with fixed part and cluster-specific effects specified only for the intercept and  $x_s$  level-1 explanatory variable.

For the class of generalized linear mixed (GLM) models marginal likelihood does not have a close form solution therefore approximation methods are applied. Generally, it is proposed to evaluate an integral  $L(\gamma, \sigma_{u_0}^2 | y, x, z)$  for the binary response model using numerical simple quadrature (integration). The more elaborate approach is adaptive quadrature that corrects weights and quadrature locations to the data for the individual clusters.

In the case of random-intercept model marginal likelihood function could be viewed as joint probability of all responses for all clusters with assumption that clusters are independent. Marginal joint probability conditional on covariates is obtained by integrating out the random intercept  $u_{j,0}$ , where  $f(u_{j,0})$  represents normal density of  $u_{j,0}$  with mean zero and variance  $\sigma_{u_j}^2$ :

$$L(\gamma, \sigma_{u_0}^2 | y, x, z) = \prod_j \int_{-\infty}^{+\infty} \prod_i g(y_{ij} | x_{ij}, z_j, u_{j,0}) f(u_{j,0}) du_{j,0}, \text{ where}$$

$$g(y_{ij} | x_{ij}, z_j, u_{j,0}) = \mu_{ij}^{y_{ij}} (1 - \mu_{ij})^{1-y_{ij}}$$

$$\mu_{ij} = 1 - F\left(-\left\{\gamma_{00} + \sum_{k=1}^K \gamma_k x_{ij} + \sum_{m=1}^M \theta_m z_{j,m} + u_{j,0} + \sum_{s=1}^S u^{(2)}_{sj} x_{ij,k}\right\}\right)$$

$$f(u_{j,0}) = \frac{1}{\sigma_{u_j,0} \sqrt{2\pi}} \exp\left(-\frac{u_{j,0}^2}{2\sigma_{u_0}^2}\right)$$

## IV. Empirical results

### 4.1. Logistic function

We begin by fitting a simple logit model for pooled data. It is done for the illustrative purpose and in order to compare the single-level structure with multilevel. In table 2 results are shown for the five most relevant for credit scoring explanatory variables. The whole set of estimated coefficients and receiver operating characteristics curve (ROC) is available in the Appendix 1.

Table 2  
Logistic regression for probability of default

$x_{ij}$	Coefficient	SE (p-value)
<u>Fixed Part</u>		
Total Income	-0.022	0.003***
Number of dependents	0.137	0.029***
Trade accounts	-0.157	0.016***
Bank advance	-1.221	0.230***
Derogatory Reports	0.203	0.074***

Logistic regression  $Pr(y_i = 1 | x) = \text{Logit}^{-1}(\beta_0 + \sum_{k=1}^K \beta_k * x_{ik})$  estimate applicant  $i$  probability of default given explanatory variables  $x_i$ .  
\*\*\* 1% significance level, \*\* 5% significance Level

### 4.2. Random-intercept and random-coefficients models

We expand the previous model and include a random-intercept at the level-2. In the nested structure all customers are clustered into 81 groups according to similarities in the socio-demographic and economic conditions of their living environments. This helps to relax the main assumption of the logit model of the conditional independence among responses for the same cluster given the covariates. The two-level structure with a varying intercept and four explanatory variables is fitted.

The random intercept represents the aggregate unobserved microenvironment specific-effect that was omitted in the baseline logit model. To explain this variability we can think that customers facing contrary environments (like rich/middle/poor areas, well-developed/less-developed facilities infrastructure) have exposure to different risks given all personal factors equal (like personal income, age, etc.).

The simple two-level model for the probability of default on income, number of dependents, ownership indicator and number of enquiries is:

$$\Pr(y_{ij} = 1 | x_{ij}, u_{j,0}) = \text{Logit}^{-1}(\alpha_j + \gamma_1 \text{Income} + \gamma_2 \text{Dependents} + \gamma_3 \text{Own}_{(rent)} + \gamma_4 \text{Enquiries})$$

It assumed that the random part follows Normal distribution with zero mean and variance  $\sigma_{u_j}^2$ :

$$\alpha_j = \gamma_0 + u_{j,0} \quad \text{random intercept for microenvironment } j = \overline{1..81}$$

$$\text{Level - 2: } u_{j,o} | x_{ijk} \sim N(0, \sigma_{u_j}^2) \quad \text{with } u_{j,o} - \text{independent across clusters } j$$

In a very simple case if all explanatory variables are equal to zero ( $x_{ij} = 0$ ) cluster error component  $u_{j,0}$  is viewed as particular living environment effect which is relevant not only for the credit worthiness assessment but also for building a market strategy given particular audience structure. In mathematical terms, varying intercept  $u_{j,0}$  for the cluster  $j$  explains additional increase or decrease in the predicted probability of default for applicants sampled from cluster  $j$  over and above population average value  $\gamma_0$  and holding all other factors fixed.

The estimation is done in STATA 10.1 by specifying Gauss-Hermite adaptive quadrature with 20 integrating points. The results for the set of main parameters are presented in the Table 3 :

Table 3

Two-level random-intercept model

	Coefficient	SE (p-value)
<u>Fixed Part</u>		
Total Income	-0.019	0.003***
Bureau Enquiries	0.225	0.023 ***
Number of dependents	0.134	0.036***
Own/Rent	-0.124	0.100 (0.220)
<u>Random part</u>		
		$\Omega_u = [16.36(4.33)]$
$\alpha_j = -5.03 (0.63) + u_{j,0}$	$[u_{j,0}^{(2)}] \sim N(0, \Omega_u)$	95% CI = [2.89; -12.95]
		Probability CI = (0.0000023%; 94%)

\*\*\* 1% significance level, \*\* 5% significance level

Importantly, the 95% confidence interval for the random intercept describes that we expect 95% of realizations of the random intercept lie within the range [2.89;-12.95] . It should not be confused with traditional interpretation of a confidence interval as an interval estimate of a population parameter<sup>5</sup>.

#### 4.3 Random-coefficients model

Regression coefficients for the pooled data give us sample estimate of population parameters and fixed for all applicants in the data set. But if we think the slope for predictor  $x_{ij}$  varies across clusters and predicts different probabilities of default we can include it on the random basis. This allows to specify a probability distribution for the slope parameter  $\beta_x$  and identify a model it follows. The estimated mean and variance of a random coefficient helps to make inference about the infinite population of microenvironments people live in and this is our main target.

In more limited sense, if we were concerned about estimates for the particular sample we would use fixed-effects approach and include dummies for each cluster in data set. The limitation of this

<sup>5</sup> See [12] Rabe-Hesketh, S., Skrondal, A. Multilevel and Longitudinal Modeling using STATA. (2008)

method is that it cannot be used in forecasting probability of default for a new cohort of applicants in a new group.

To elaborate the previous structure we include more predictors and specify random slopes for the covariates  $x_{enquiries}$  (number of Bureau enquiries) and  $x_{DR}$  (derogatory reports). Random-coefficients are also viewed as the interaction effect between individual level covariate  $x_{ij}$  and cluster-level indicator and show variation in covariates' predictive magnitude across microenvironments.

The two-stage formulation for the multilevel model with random slopes is following:

$$\Pr(y_{ij} = 1 | x_{ij}, u_{j,0}) = \text{Logit}^{-1}(\gamma_0 + \gamma_1 \text{Income} + \gamma_2 \text{Dependents} + \gamma_3 \text{Own}_{(rent)} + \dots + \beta_j^{enq} \cdot \text{Enquiries} + \beta_j^{DR} \cdot \text{DR})$$

$$\beta_j^{enq} = \gamma_{enq} + u_{j,enq}$$

$$\beta_j^{DR} = \gamma_{DR} + u_{j,DR} \quad , \quad u_{j,enq}, u_{j,DR} | x_{ijk} \sim N\left(0, \Sigma_u = \begin{bmatrix} \sigma_{enq}^2 & \sigma_{enq,DR} \\ \sigma_{DR,enq} & \sigma_{DR}^2 \end{bmatrix}\right)$$

Given  $x_{ij}$  random coefficients at level-2 follow a bivariate normal distribution with variance-covariance matrix  $\Sigma_u$  and population average slopes  $\gamma_{enq}, \gamma_{DR}$ .

The maximum likelihood estimates for the random and fixed part of the scoring model are presented in the table below:<sup>6</sup>

Table 4

Two-level random-coefficients model

$x_{ij}$	Coefficient	SE (p-value)
<b>Fixed Part</b>		
Total Income	-0.025	0.0029***
Number of dependents	0.101	0.029***
Own/Rent	-0.141	0.100 (0.159)
Bank advance	-0.476	0.078***
Age	-0.024	0.003 ***
Trade accounts	-0.242	0.022***
...		
<b>Random part</b>		
N° of enquiries	$\beta_j^{enq} = 0.125(0.026) + u_{j,enq}$	$\begin{bmatrix} u_j^{enq} \\ u_j^{DR} \end{bmatrix} \sim MVN(0, \Sigma_u) \quad \Sigma_u = \begin{bmatrix} 1.29 (0.63) & \dots \\ 2.36 (0.92) & 4.94 (1.81) \end{bmatrix}$
Major DR	$\beta_j^{DR} = 0.172 (0.065) + u_{j,DR}$	

\*\*\* 1% , \*\* 5% -significance level

The estimated regression fixed effects (upper part of the Table 4) agree with Logit and a Random-intercept model. As it might be expected, default probability decreases with a higher total income, more experience in credit and debit banking products use, number of trade accounts and the ownership

<sup>6</sup> Results obtained from STATA 10 and MLwin (MQL-2) are very similar, except that MQL-2 estimates a little bit downward biased.

indicator. It should be mentioned that with nonlinear models we cannot directly interpret obtained values for coefficients as marginal effects like in linear model because first derivatives have a nonlinear functional form in generalized linear models. The rough interpretation of a slope magnitude may be obtained by simply dividing its estimated value by 4. For example, population average estimate for the random coefficient  $x_{\text{derogatory reports}}$  raises probability of default by 4.3% if number of major derogatory reports increases by 1.

The estimated level-2 variances for the varying slope for the  $x_{\text{enquiries}}$  equal  $\sigma_{\text{enq}}^2 = 1.29$  and for the  $x_{\text{derogatory reports}}$   $\sigma_{\text{DR}}^2 = 4.94$  on the logit scale illustrate residual variability across groups in predicting default over and above population slopes  $\gamma_{\text{enq}}, \gamma_{\text{DR}}$ .

Generally, we could think there are more customers with good credit histories in the high income areas with high percentage of families own a house, developed facilities infrastructure and low unemployment rate. Therefore given 1 unit increase in number of Bureau enquiries, from  $x_{ij}^{\text{enq}} = n$  to  $(n + 1)$  lead to a lower increase in probability of default for the applicant sampled from these microenvironments. Subsequently, for microenvironments with low average income, not-developed facilities infrastructure and high unemployment we face more customers with bad credit histories meaning they frequently apply for a loan and often are rejected, have many accounts past due or delinquencies. The illustration of random-slope effect is better to represent graphically and that is done in the section V ( see graph 5.a, 5.b).

#### 4.4 Monte Carlo Markov chain estimation for mixed-effects model

Some extensions are applied to the last model and complementary to the previous structure number of dependents and house ownership indicator are included as additional random coefficients. The variance-covariance matrix is constrained to have independent structure – all covariances are set to zero. That would speed up the estimation process as number of parameters to evaluate is noticeably decreased.

**Table 5**

Two-level mixed-effects model: Monte Carlo Markov chain estimation

	Coefficient	SE (p-value)
<u>Fixed Part</u>		
Total Income	-0.024	0.003
Bank advance	-0.414	0.079
Age	-0.027	0.004
Number of dependents	-1.203	0.198
Past due	0.284	0.037
Trade accounts	-0.264	0.026
.... (others, not listed here)	...	
<u>Random part</u>		
N° of enquiries: $\beta^{enq}_j = 0.130 (0.023) + u_j^{enq}$	$\begin{bmatrix} u_j^{enq} \\ u_j^{dep} \\ u_j^{DR} \\ u_j^{Own} \end{bmatrix} \sim N(0, \Sigma_u)$	
N° of dependents: $\beta^{dep}_j = -1.203 (0.198) + u_j^{dep}$		
Minor DR: $\beta^{DR}_j = 0 + u_j^{DR}$		
Own/ Rent: $\beta^{Own}_j = 0 + u_j^{Own}$		
	$\Sigma_u = \begin{bmatrix} 2.305(0.758) & 0 & 0 & 0 \\ 0 & 1.781(0.581) & 0 & 0 \\ 0 & 0 & 0.365(0.158) & 0 \\ 0 & 0 & 0 & 1.027(0.487) \end{bmatrix}$	

One of the most critical points remained for a whole range of the fitted models is that the coefficient for the ownership indicator  $x_{Own(rent)}$  appeared to be only weakly significant at 10% level or even less as in the case of estimates represented in Table 4 . This is a curious case because typically credit managers would determine this covariate as a very important predictor and additionally with other personal information decide whether to accept or reject a customer. One possible answer for this confusing result is that the ownership indicator is indeed a significant variable but there is no fixed for all applicants coefficient but rather a specific random effect that accounts for unobserved heterogeneity across microenvironments. In other words, we suppose that a population-average coefficient for ownership indicator is zero. The estimated variance of a random-slope is obtained to be 1.027 with standard error - 0.48.

The varying-coefficient  $\beta^{dep}_j$  for covariate the  $x_{number\ of\ dependents}$  explains the variability of slope across environments with different economic and demographic conditions. Some individuals become more responsible given they have more persons in the family to take care of . This is especially true for the applicants with high level of education and high life standards (like the academic society). On the other hand, there are customers for whom more dependents in the family indicate higher exposure to risk and as a consequence higher probability of failure.

## V. Predicted Probabilities of Default

The main idea behind a multilevel structure for the scoring model is to produce a better fit to the data and obtain more accurate estimates for the predicted probabilities given set of explanatory variables. In order to compare predictive quality of single-level and multilevel regressions we apply the Brier score.

Brier score helps to evaluate how large is the average squared deviation of the predicted probabilities from the actual outcomes, a lower score represents higher accuracy. It could also be interpreted as the per observation squared error produced by the fitted model:

$$Brier\ Score = \frac{1}{N} \sum_1^N (\theta_{ij} - \widehat{p}_{ij})^2, \text{ where } \theta_{ij} = \begin{cases} 1 & \text{actually observed default} \\ 0 & \text{non - default} \end{cases}$$

The results for Brier scores are displayed in the Table 6. If we compare scores for the set of mixed-effects models with single-level model it is clear that the logistic regression produces crudest estimates for the probabilities of failure on a loan.

**Table 6**

Predictive accuracy of Logit and Multilevel models

Model	Brier Score
Logistic Regression	0.08714
Random-intercept model	0.07385
Random-coefficients model	0.07087
Random-coefficients model with Monte Carlo Markov chain estimation	0.05652

More transparently probability estimates could be represented using graphs. Therefore we illustrate average predicted probabilities for the pooled data model (logistic regression) and a two-level random-effects model in the Figure 2.

Evidently, predicted probabilities are heterogeneous across microenvironments and the logistic regression fails to account for cluster-specific effects. Estimations obtained with Monte Carlo Markov chain approach shows more accurate results and the red curve is closer to the actually observed default probabilities.

Figure 2

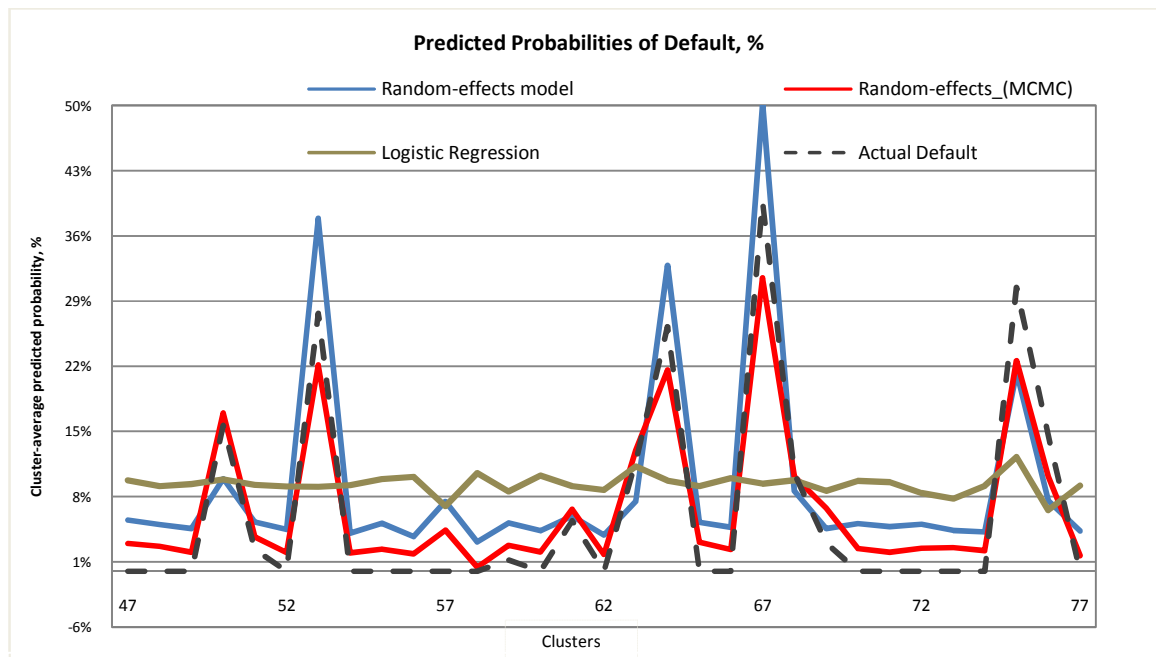
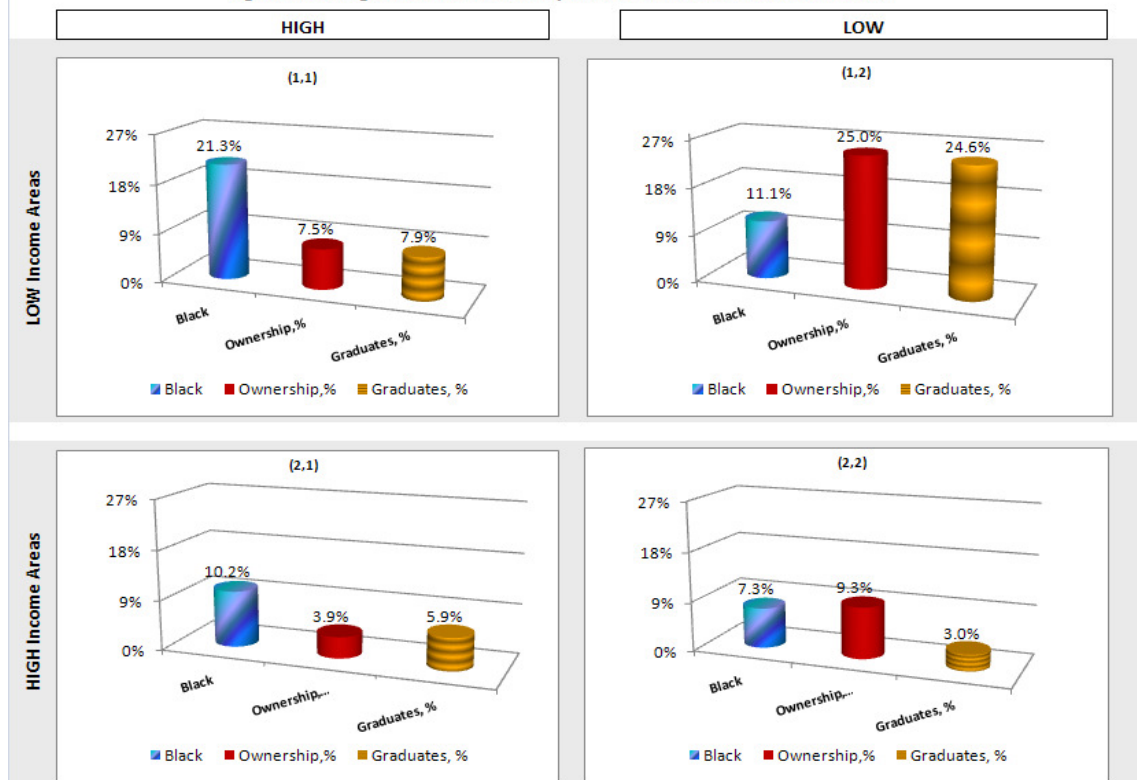


Figure 2. Prediction after estimation for Logit model and Two-level mixed-effects models for different microenvironments.

It is commonly believed that income is the main determinant in a decision making process for a credit issuance. That is quiet correct and this information gives a very good assessment of a customer financial solvency, but does it play the same tremendous role in predicting a default behavior. Probably no, because we also need some information about other triggering default factors like living environment that describes economic and demographic conditions of customer’s surroundings. For example, two applicants from to the same high income category but reside within dissimilar micro-socio-environments are influenced by different risk factors. Typically, it is matter whether person lives in rich or middle income area with low unemployment, high percentage of house ownership in the district, low crime rates and a well-developed retail stores infrastructure or his surroundings are reverse given the same level of personal assets ( like home ownership, earnings and etc.). This is also the answer to the question how relevant is market and auxiliary data in probability forecasting with the multilevel modeling<sup>7</sup>. To illustrate this idea we plot the average predicted probabilities for the customers sampled from contrary microenvironments in the Figure 3. The charts are represented in the matrix form. Rows illustrate high and low income areas and columns show high (low) percentage of house ownership in the district, percentage of African-American residents and presence of college graduates in the area job market.

<sup>7</sup> Another important issue is occupation field and stage of employment cycle that could provide useful information of applicant regular duties and responsibilities, self-discipline and what kind of social network he belongs to. We do not consider this classification scheme in this paper because of lack of initial data.

Figure 3 . Average Predicted Probability of Default for the microenvironments



(1,1) Average probability of default for the microenvironments with low average income and high density of black population, high percentage of house ownership and many college graduates .

(2,1) Average estimated probability of default for the microenvironments with high average income and high density of black population, high percentage of house ownership and many college graduates.

(1,2) Average estimated probability of default for the microenvironments with low area income and low density of black population, low percentage of house ownership and low percentage of college graduates .

(2,2) Average probability of default for the microenvironments with high area income and high density of black population, high percentage of house ownership and many college graduates .

Previously we mentioned that the clustering was done using the following determinants : average income area, percentage of people own a house, percentage of African-American (Spanish) residents in the area, percentage of college graduates in the area job market, degree of infrastructure<sup>8</sup> development and percentage of employment in the district. It is not feasible to plot probability of default against all these determinants. Therefore we take 20 % of highest and 20% of lowest income areas (that is the rows in the graph matrix) and calculate average predicted probabilities for the microenvironments with high (low) percentage of families own a house, high (low) percentage of college graduates and high(low) percentage of African-American in the district (columns). These diagrams show variability in predicted probabilities across microenvironments with different socio-demographic and market conditions.

Noticeable, estimated probability of default is more pronounced in low income areas than in high (chart 1,1 and 1,2 versus 2,1 and 2,2) given all other factors equal. The chart 1,1 (1,2) shows the

<sup>8</sup> Infrastructure is determined in terms of the basic facilities, services, and installations needed for the functioning of a community or society: medical, shopping, dining, communication facilities and etc.

representative applicant within poor districts is more likely to default on a loan with calculated probability 21,3% ( or 11,1%) versus 10,2%(or 7,3%) than in the rich or middle income areas.

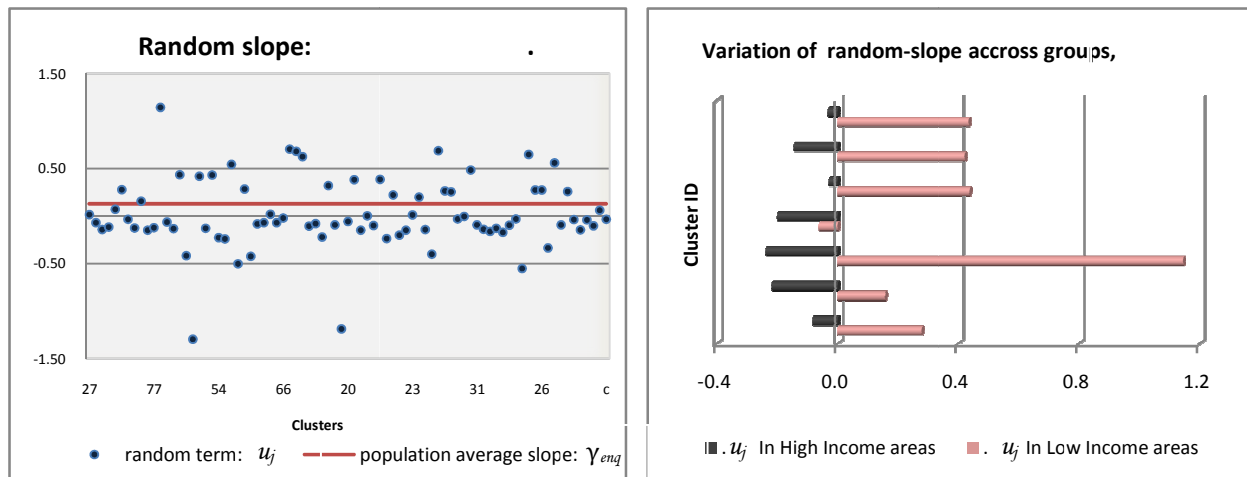
There is also evidence that within poor districts probability of default is smaller for areas with high percentage of a home ownership (7.5%), small density of African-American (Spanish) residents (11.1%) and high percentage of college graduates (7.9%). The same is true for the richer areas.

In general the unobserved microenvironment-specific effects help to explain variability in the predictions over and above population average value and lead to more a more accurate riskiness assessment of lending to a particular group of customers (Figure 4). The other relevant aspect is that the implementation of multilevel set up for the credit scoring model is to make inference of the infinite population of different micro-socio-environments with various conditions.

On the next figure we illustrate the environment-specific effects estimates for the random slope for the covariate:  $X_{Enq}$  - number of Credit Bureau enquiries (shows how many times applicant's credit file was requested by other financial institutions).

Figure. 4

Estimated microenvironment-specific effect in the random-coefficient multilevel model



4.a) Estimates for the random slope for the covariate - number of Bureau enquiries.

4.b) Microenvironment effect for applicants with high per capita income and large number of enquiries sampled from areas with high(low) average income.

As defined in the section 4.3 the random-coefficient  $\beta_{enq} = \gamma_{enq} + u_j$  consists of population average coefficient  $\gamma_{enq} = 0.130$  and random term  $u_j$  which varies across microenvironments,  $j=1..81$ . Figure 4.a shows that the cluster-specific effect  $u_j$  explains change in predictive magnitude of the explanatory variable  $X_{Enq}$  over and above population average value 0.130. This graphical interpretation extends the idea of applicants with good and bad credit histories discussed in the section 4.3.

For a more coherent representation we can plot microenvironment-specific effect  $u_j$  for the top 1000 of applicants for a loan with large number of enquiries and high personal income. The 35% of this subsample belongs to the low income areas and 40% to the high.

Generally, poor districts have higher share of applicants with bad credit histories and large number of credit file searches states that they often apply for a loan and many times are rejected or have many accounts past due. Oppositely, in richer areas we mainly face customers with good credit histories meaning that they frequently apply for a credit and have pay it back without delinquencies. Accordingly, in the first case the microenvironment effects are estimated to be smaller than zero and predict smaller increase in probability of default (black lines, Figure 4.b). Oppositely, in the second case cluster-specific residuals are positive and concerned with higher exposure to failure (red lines, Figure 4.b).

## CONCLUSION

By considering a multilevel hierarchical model for application credit scoring we would like to provide some insight into the process of probability of default prediction from the microenvironment-specific perspectives. The multilevel set up for a scorecard helps to make inference of the infinite population of different environments with various economic and demographic conditions. This also helps to account for the unobserved heterogeneity across clusters and solve the omitted variables problem. In general the primary benefit of the improved efficiency of a scorecard is higher accuracy of the model prediction.

The mixed-effects model treats applicants for a credit as level-1 units which are nested within micro-socio-environments, level-2 subjects. This multilevel nested structure allows specifying random effects at the higher levels and calculate between and within clusters variation.

The hierarchical clustering of applicants was done using the following determinants : average income area, percentage of house ownership in the district, density of African-American (Spanish) population, percentage of college graduates in the area job market, degree of infrastructure development and unemployment rate. The defined micro-socio-environments account for similarities in economic and socio-demographic conditions in the customers' living surroundings.

The model was built using two-stage formulation by the extending general logistic regression and allowing intercept and slopes to vary across groups. In the case of random-intercept model cluster-specific intercept  $\alpha_j$  represents the aggregated unobserved effect that was omitted in the baseline logit model. Customers reside in dissimilar environments have exposure to different risk factors and it is relevant to take this pattern into account in the probability of default forecasting. There is also evidence that within higher income areas there is a larger amount of customers with good credit histories meaning that they frequently apply for a credit and pay it back without delinquencies.

## References

- [1] Goldstein H. (2003). *Multilevel statistical models*. London: Arnold, 3rd edition.
- [2] Goldstein H. , Rasbash J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, 159:505-513.
- [3] Goldstein H., Rasbash J. , Browne W.J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, 1:223-231.
- [4] Anderson R. (2008). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press.
- [5] Browne W.J. (1998). *Applying MCMC methods to multilevel models*. Ph.D. thesis, University of Bath.
- [6] Guang Guo, Hongxin Zhao (2000). Multilevel modeling for binary data. *Annu.Rev.Sociol.* # 26 pp 441-462.
- [7] Bryk A.S. , Raudenbush S.W. (1992). *Hierarchical linear models*. Newbury Park, California: Sage.
- [8] McCullagh P. , Nelder J. (1989). *Generalized linear models*. London: Chapman & Hall.
- [9] Gelman A., Brown C.H., Carlin J.B., Wolfe R.(2001). *A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes*. Oxford University Press.
- [10] Rodriguez G., Elo I. (2003). Intra-class correlation in random-effects models for binary data. *The Stata Journal*, 3, Number 1, pp.32-46.
- [11] Greene W. (1992). *A statistical model for credit scoring*. Working paper. Stern School of Business.
- [12] Gelman A., Hill J. (2007). *Data analysis using regression and multilevel /hierarchical models*. Cambridge University Press.
- [13] Rabe-Hesketh S., Skrondal A. (2008). *Multilevel and Longitudinal Modeling using STATA*.
- [14] Rabe-Hesketh S., Skrondal A., Pickles A. *Generalized multilevel structural equation modeling*. *Psychometrika*.
- [15] Rabe-Hesketh S., Skrondal A., Pickles A. (2001). Generalized multilevel parameterization of multivariate random effects models for categorical data. *Biometrics*, 57, 1256–1264.
- [16] Raudenbush S., Yang M. (1988). *Maximum likelihood for hierarchical models via high order Multivariate Laplace approximation*. University of Michigan.
- [17] Raudenbush S., Bryk A., Cheong Y. F., Congdon (2000). *HLM 5. Hierarchical linear and non linear models*. Lincolnwood, IL: Scientific Software International, Inc.
- [18] Anderson D. A., Aitkin M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society B*, 47, 203–210.
- [19] Stiratelli R., Laird N., Ware J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics* 40, 961–971.
- [20] Zeger S. L., Karim, R. M. (1991). Generalized linear models with random effects: a Gibbs sampler approach. *Journal of the American Statistical Association*, 86, 79–86.
- [21] Cameron C.A., Trivedi P.K.(2005) *Microeconometrics: Methods and Applications*. Cambridge University Press, NY.

## Appendix 1

### Logistic Model.

Results for the logit regression are obtained using STATA 10 and represented in the following STATA format table :

```

Logistic regression                               Number of obs   =    10420
                                                    LR chi2(19)    =    585.35
Log likelihood = -2978.9704                       Prob > chi2     =    0.0000
  
```

default	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Total_income	-.0166411	.0026438	-6.29	0.000	-.0218228 -.0114594
Bureau_enqr	.2371897	.0175745	13.50	0.000	.2027442 .2716351
Bank_advance	-.4003162	.0733416	-5.46	0.000	-.5440631 -.2565692
Addit_card	.1949542	.0877622	2.22	0.026	.0229435 .3669649
Major_DR	.2771034	.0715169	3.87	0.000	.1369328 .417274
Minor_DR	.3052774	.0519774	5.87	0.000	.2034035 .4071512
Trade accts	-.0231419	.0090029	-2.57	0.010	-.0407874 -.0054964
Age	-.0131217	.0046028	-2.85	0.004	-.0221429 -.0041004
Dependents	.0947398	.0298789	3.17	0.002	.0361782 .1533013
Employed_mth	.0016611	.000599	2.77	0.006	.000487 .0028352
Professional	-.4360813	.1266121	-3.44	0.001	-.6842365 -.1879262
Management	-.1770532	.1382986	-1.28	0.200	-.4481134 .0940071
Military	.5645593	.2003156	2.82	0.005	.171948 .9571707
Clerical	.2305317	.1157938	1.99	0.046	.0035801 .4574834
Sales	-.0454535	.1318155	-0.34	0.730	-.3038072 .2129001
Own_Rent	-.095	.0812471	-1.17	0.242	-.2542413 .0642414
Cur_adr_mth	.0000888	.0006176	0.14	0.886	-.0011217 .0012992
Pr_adr_mth	.0007628	.0004149	1.84	0.066	-.0000503 .0015759
Prev_cardhld	-.3704309	.1677541	-2.21	0.027	-.6992228 -.0416389
Act_accnt	-.2219946	.0260223	-8.53	0.000	-.2729974 -.1709918
Avr_rev_cr	.0128865	.0036941	3.49	0.000	.0056463 .0201268
Past_due	.1965916	.0281697	6.98	0.000	.14138 .2518031
Constant	-1.25247	.159446	-7.86	0.000	-1.564978 -.9399611

### Receiver-operating Characteristics Curve for the Logit model:

