Impact of social factors on microloans defaults

Cedric Huk Abekah Koffi [1], Abigail Baidoo [2],
Olivier Menoukeu Pamen [3],
Viani Biatat Djeundje[4],

June 13, 2023

UNIVERSITY OF LIVERPOOL

Table of Contents

The Framework

Microfinance

- Institutions that provide in small loan amounts, and other financial services to low-income individuals
- They are mostly established in developing countries
- Target is small low-income individuals, small-scale businesses
- Low individuals especially carry high risk of default
- Microfinance expose themselves to high risk of bankruptcy, and hence tend to charge astronomical interest risk to counteract such risk


SINCE 1857

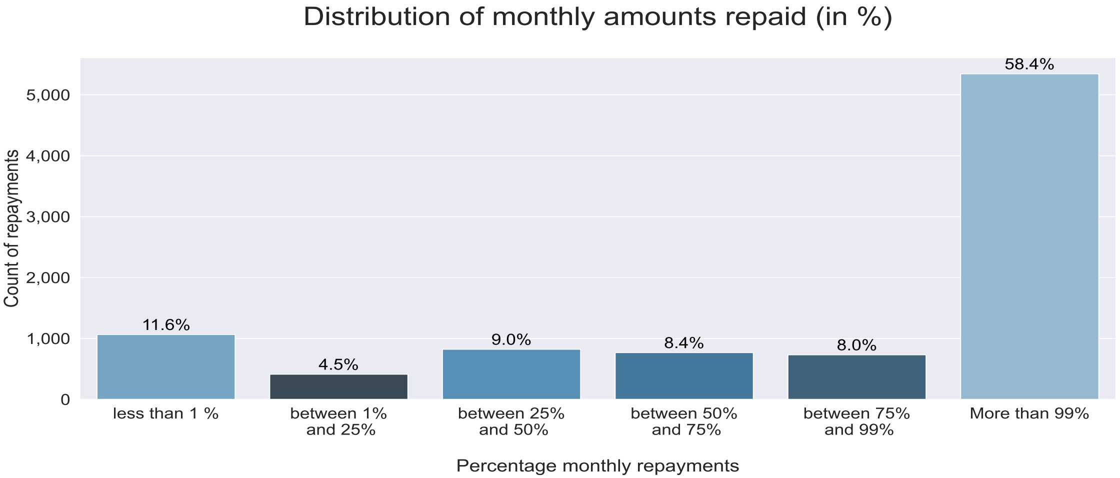
Problems and objectives

- Traditional modeling of loan defaults looks at whether someone will default at term of repayments; we focus on repayments dynamics throughout the loan duration [1]
- Identify at the local levels factors which may have higher impact on loan delinquency; i.e. that standard models do not take into consideration
- Minimize the number of non performing loans (NPLs) since microfinance institution are heavily regulated by central banks in this regard - which is directly related to reducing loan delinquency
- Estimate with an acceptable level of accuracy transition probabilities from state i at time $t-1$ to state j at time t.

The data

- $8,303$ observation
- $1,716$ customers
- Covariates information
  - Loan information : Principal, interest, duration of repayment, frequency of repayment, group or individual, branch, type of loan, balance
  - macroeconomic variables (lagged) : consumer price index, foreign exchange rate, Bank of Ghana lending rate, inflation
  - *Social variables* : pertaining to the Ghana's socio-economic settings
  - Demographics : Age of customer, marital status, gender

Some insights from the data



Our definition of delinquency

- Account i is in delinquency when the cumulative amount repaid by this account at time t less than 82% of the *cumulated* agreed amount to repay at such time t; in this situation we consider a 2 state model
- In a more dynamic setting where we do not look at cumulative repayments, we define 2 states which define the level of delinquency of account i. Let's consider $A(t)$ to be the amount account i has to repay at time t, and $x_i(t)$ to be the amount account repaid at time t by this account, then
  - Account i is in state 3 if $0 \leq x_i(t) < 0.5A(t)$
  - Account i is in state 2 if $0.5A(t) \leq x_i(t) \leq 0.9A(t)$
  - Account i is in state 1 if $x_i(t) > 0.9A(t)$
- We consider no absorbing state

The fixed effect model

Setup of the model

- Consider a portfolio of accounts i associated with the process $\boldsymbol{y}_i = \{y_{i,hj}(t),\ t \geq 0\}_{(h,j)\in\mathcal{S}}$, $\mathcal{S} = \{(h,j)\ ,\ h \neq j\}$
- $\mathcal{S}$ is the set of all possible transition-types $(h,j), h \neq j$.
- We assume that $y_{i,hj}(t)$ follows a Bernoulli distribution and is defined as

$$y_{i,hj}(t) = \begin{cases} 1 & \text{if account } i \text{ in state } j \text{ at time } t \mid \text{account } i \text{ was in state } h \text{ at time } t-1, \\ 0 & \text{if account } i \text{ in state } h \text{ at time } t \mid \text{account } i \text{ was in state } h \text{ at time } t-1. \end{cases}$$

- For cases where an account i makes a transition $h \to j^*$, $j^* \notin \{h,j\}$, we assume the process is interval-censored and non-information [2, 1]

Time-dependent transition probabilities

- We model the transition probabilities directly using the logit link function

$$q(f(x)) = 1/\left(1 + e^{-f(x)}\right)$$

- The time dependent transition probability is then given as

$$\begin{cases} \mathbb{P}\left(y_{i,hj}(t) = 1\right) = q_{i,hj}(t) \\ \mathbb{P}\left(y_{i,hj}(t) = 0\right) = 1 - q_{i,hj}(t) \end{cases} \quad (1)$$

Time-dependent transition probabilities (cont'd)

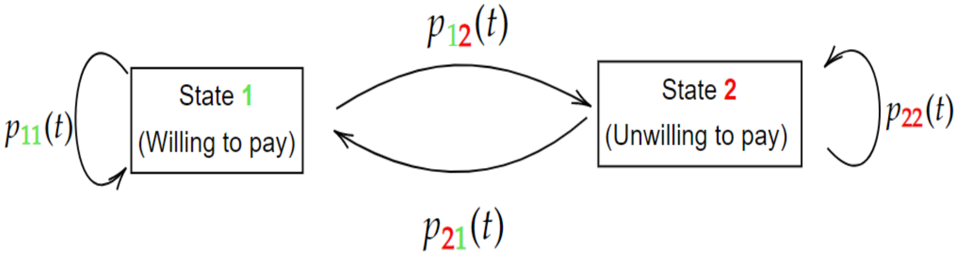
- More specifically, $q_{i,ht}(t)$ is time-dependent and is defined as

$$q_{i,hj}(t) = \frac{1}{1 + \exp(\alpha_{hj}(t) + \boldsymbol{\beta}_{hj}(t)^T X_{i,hj}(t))}, \quad (2)$$

  where

  - $\boldsymbol{\beta}_{hj}$ is a vector of fixed-effect coefficients to estimate,
  - $\alpha_{hj}(t) = \sum_{r=1}^{c} B_r(t)\varphi_{hj,r}$, where is a B_r is a B-spline basis function at time t,
  - $\boldsymbol{\varphi}_{hj} = (\varphi_{hj,1}, ..., \varphi_{hj,c})$ is a vector of B-spline coefficients to estimate,
  - $X_{i,hj}(t)$ is a vector of possibly time-dependent covariates.

Graphical representation 2 states recurrent model



Log-likelihood function and estimation parameters

- To estimate the parameters $\boldsymbol{\gamma_{hj}} = (\boldsymbol{\beta}_{hj}, \boldsymbol{\varphi}_{hj})$, we write the transition-dependent likelihood function as a product of Bernoulli PMF's stratified on event times

$$L(\gamma_{hj}) = \prod_{t \in I \subset \mathbb{N}} \prod_{i \in \mathcal{R}_{hj}(t)} q_{i,hj}(t)^{y_{i,hj}(t)} \left(1 - q_{i,hj}(t)\right)^{(1 - y_{i,hj}(t))}, \qquad (3)$$

  where $(h,j) \in \mathcal{S}$, and $\mathcal{R}_{hj}(t)$ is the set of accounts at risk of transition (h,j) before time t.

- To reduce chances of numerical overflow in the estimation of $\boldsymbol{\gamma_{hj}}$, we deal with the following log-likelihood instead

$$l(\gamma_{hj}) = \sum_{t \in I \subset \mathbb{N}} \sum_{i \in \mathcal{R}_{hj}(t)} y_{i,hj}(t) \log(q_{i,hj}(t)) + (1 - y_{i,hj}(t)) \log\left(1 - q_{i,hj}(t)\right) \qquad (4)$$

Log-likelihood function and estimation parameters (cont'd)

- The vector estimate of (3) is given by

$$\hat{\gamma}_{hj} = \arg\min_{\gamma_{hj}} \left(-l(\gamma_{hj})\right) . \tag{5}$$

- We use the efficient Python optimization library Scipy [3] to minimize (5).
- Next, our aim is account for the effect of unobserved covariates and model the possible dependence among account i's repayments

The random effects (frailties) model

Setting up the complete data

- To model for the effects of unobserved covariates, we assume that we are dealing with a incomplete data problem
- We consider the complete data vector $(\boldsymbol{y}^T, \boldsymbol{u}^T)^T$ with $\boldsymbol{y} = (\boldsymbol{y}_i)_{i\in\{1,\dots,n\}}$, and $\boldsymbol{u} = (\boldsymbol{u}_i)_{i\in\{1,\dots,n\}}$ representing the vector of frailty vectors,
- $\boldsymbol{u}_i = \left(u_{i,hj}\right)_{(h,j)\in\mathcal{S}}$ is the frailty vector associated to customer i,
- We consider $\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{u}$, $\boldsymbol{u}_i \perp\!\!\!\perp \boldsymbol{u}_j$ for $i \neq j$, as well as $\boldsymbol{u}_{i,hj} \perp\!\!\!\perp \boldsymbol{u}_{i,h^*j^*}$ for $(h,j) \neq (h^*,j^*)$, so we assume the framework of shared frailties [4] among event of type (h,j) for account i

Complete data likelihood for account i (cont'd)

- We write the time dependent transition probability as

$$q_{i,hj}(t) = \frac{1}{1 + \exp\left(\alpha_{hj}(t) + \boldsymbol{\beta}_{hj}(t)^T X_{i,hj}(t) + u_{i,hj}\right)}, \tag{6}$$

  where all similar terms are defined as in the fixed effects model.
- The new vector of parameters to estimate is $\xi_{hj} = \left(\boldsymbol{\varphi}_{hj}, \boldsymbol{\beta}_{hj}, \phi_{hj}\right)$

Complete data likelihood for account i (cont'd)

- The contribution to the joint transition-dependent likelihood from an account i at time t can be written as

$$L_i = L_{\left(y_{i,hj}(t),u_{i,hj}\right)}(\boldsymbol{\xi}_{hj}) = g_{u_{i,hj}}(\boldsymbol{\xi}_{hj}) \prod_{\substack{t\in I\subset\mathbb{N} \\ i\in\mathcal{R}_{hj}(t)}} L_{y_{i,hj}(t)|u_{i,hj}}(\boldsymbol{\xi}_{hj}) \tag{7}$$

where $L_{\left(y_{i,hj}(t)|u_{i,hj}\right)}(\boldsymbol{\xi}_{hj})$ is the pmf of the Bernoulli with $p = q_{i,hj(t)}$ and g_{u_i} is the univariate Gaussian density

$$g_{u_i}(\phi_{hj}) = g_{u_i}(\boldsymbol{\xi}_{hj}) = \frac{\exp\left(-\frac{1}{2}\frac{u_{i,hj}^2}{\phi_{hj}}\right)}{\sqrt{(2\pi\phi_{hj})}}. \tag{8}$$

Complete data likelihood for account i when $\dim(u_i) > 1$ (cont'd)

- The contribution to the joint transition-dependent likelihood from an account i at time t can be written as

$$L_i = L_{(y_i,u_i)}(\boldsymbol{\xi}) = g_{u_i}(\boldsymbol{\xi}) \prod_{(h,j)\in\mathcal{S}} \prod_{\substack{t\in I\subset\mathbb{N} \\ i\in\mathcal{R}_{hj}(t)}} L_{y_{i,hj}(t)|u_{i,hj}}(\boldsymbol{\xi}_{hj}) \tag{9}$$

where $L_{\left(y_{i,hj}|u_{i,hj}\right)}(\boldsymbol{\xi}_{hj})$ is defined as before and g_{u_i} is the multivariate Gaussian density with diagonal covariance matrix.

The complete data likelihood

- The contribution of each account to the final transition-depend log-likelihood can be expressed as

$$l(\boldsymbol{\xi}_{hj}) = \sum_i \log(L_i(\boldsymbol{\xi}_{hj})) \tag{10}$$

Estimation of parameters

- $\boldsymbol{\xi}_{hj}$ is estimated by integrating out the effects $\boldsymbol{u}$ from (10), i.e.

$$\mathbb{E}_{\boldsymbol{U}|\boldsymbol{\xi_{hj}}}\left[l\left(\boldsymbol{\xi}_{hj} \mid \boldsymbol{y}, \boldsymbol{u}\right)\right] = \int_{\mathbb{R}^n} l\left(\boldsymbol{\xi}_{hj} \mid \boldsymbol{y}, \boldsymbol{u}\right) g_{\boldsymbol{U}|\boldsymbol{\xi}_{hj}}(\phi_{n\times n}) d\boldsymbol{u} \quad (11)$$

  where n is the number of accounts, $\phi_{n\times n}$ is a diagonal covariance matrix, and $g_{\boldsymbol{U}|\boldsymbol{\xi}_{hj}}$ is the multivariate Gaussian conditional density on $\boldsymbol{\xi}_{hj}$.
- (11) is not available in closed form, so we need quadrature techniques [3] or Monte Carlo techniques [5].
- $\hat{\boldsymbol{\xi}}_{hj}$ can then be estimated by minimizing

$$\arg\min_{\boldsymbol{\xi}_{hj}}(-\mathbb{E}_{\boldsymbol{U}|\boldsymbol{\xi_{hj}}}\left[l\left(\boldsymbol{\xi}_{hj} \mid \boldsymbol{y}, \boldsymbol{u}\right)\right]), \quad (12)$$

Goodness of fit

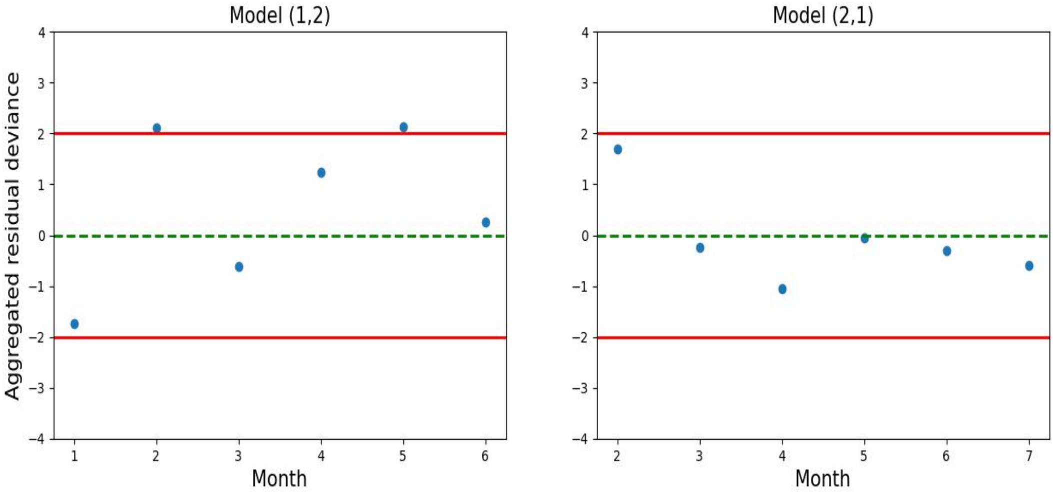
Aggregated deviance residuals

- Since predictions are made monthly and are dependent on the risk set $\mathcal{R}_{hj}(t)$, we aggregate the deviance residuals accordingly and define he deviance residual at time t as

$$D_{hj}(t) = sign(O_{hj}(t) - E_{hj}(t)) \left(2 \left(O_{hj}(t) \log \left(\frac{O_{hj}(t)}{E_{hj}(t)}\right) + (N_{hj}(t) - O_{hj}(t)) \log \left(\frac{N_{hj}(t) - O_{hj}(t)}{N_{hj}(t) - E_{hj}(t)}\right)\right)\right)^{0.5}$$

  where $O_{hj}(t)$ and $E_{hj}(t) = \sum_{i \in \mathcal{R}_{hj}(t)} \hat{q}_{i,hj}(t)$ are the total number of observed transitions and total predicted number of transitions from state h at time $t-1$ to state j at time t respectively.

Goodness of fit of fixed-effects model (cont'd)



Statistical significance of parameters in fixed-effects model

p-values computed based on $2,000$ resamples of training data

Covariates $(1,2)$	p-value $(1,2)$
Main branch	0.0
Age	0.0
Lagged CPI	0.0
Lagged FX	0.0
Lagged OI	0.009018
Long vacation	0.0
Eid	0.259519
Gender	0.145291
Group	0.0
Monthly	0.01002
Married	0.001002
Interest rate	0.0
Cub. Spline coef. 1	0.0
Cub. Spline coef. 2	0.403808
Cub. Spline coef. 3	0.0

Covariates $(2,1)$	p-value $(2,1)$
Main branch	0.828657
Age	0.002004
Lagged CPI	0.674349
Lagged FX	0.0
Lagged OI	0.019038
Long vacation	0.599198
Eid	0.0
Gender	0.213427
Group	0.343687
Monthly	0.002004
Married	0.189379
Interest rate	0.001002
Cub. Spline coef. 1	0.0
Cub. Spline coef. 2	0.003006
Cub. Spline coef. 3	0.0

Predictions

Accuracy of predictions for fixed effect model

- For accuracy we rely on the cumulative matrix

$$P(t_1, t_2) = \prod_{t=t_1+1}^{t_2} P(t) \tag{13}$$

- The model yields on average an accuracy of 60% or more

Impact of the frailties on understanding customers behaviour

Effect of frailties on an account i

Covariates	Estimate
Main branch	0.352360
Age	1.651142
Lagged CPI	-6.569061
Lagged FX	-1.704136
Lagged OI	3.873440
Long vacation	3.838679
Eid	2.336820
Gender	0.818518
Group	0.653378
Monthly	0.848854
Married	-0.159340
interest rate	1.238705
Cub. Spline coef. 1	0.167450
Cub. Spline coef. 2	0.011688
Cub. Spline coef. 3	0.887709
ϕ_{12}	0.223239
ϕ_{21}	0.047906

What's next?

- Model the effects of time-dependent frailties on delinquency
- Model the optimal interest rates the company should assign to customer i at loan disbursement
- Estimate the probability of default of loan groups under dependency settings
- Application of our models to company data to set up and manage their loan portfolio over a period of time as case study.

Thank you

References I

Viani Biatat Djeundje and Jonathan Crook.
Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards.
European Journal of Operational Research, 271(2):697–709, 2018.

Zhigang Zhang and Jianguo Sun.
Interval censoring.
Statistical methods in medical research, 19(1):53–70, 2010.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al.
Scipy 1.0: fundamental algorithms for scientific computing in python.
Nature methods, 17(3):261–272, 2020.

References II

Andreas Wienke.
Frailty models in survival analysis.
CRC press, 2010.

Geoffrey J McLachlan and Thriyambakam Krishnan.
The EM algorithm and extensions.
John Wiley & Sons, 2007.

Michael Roimi, Rom Gutman, Jonathan Somer, Asaf Ben Arie, Ido Calman, Yaron Bar-Lavie, Udi Gelbshtein, Sigal Liverant-Taub, Arnona Ziv, Danny Eytan, et al.
Development and validation of a machine learning model predicting illness trajectory and hospital utilization of covid-19 patients: a nationwide study.

References III

Journal of the American Medical Informatics Association, 28(6):1188–1196, 2021.

Terry M Therneau and Patricia M Grambsch.
Modeling survival data: extending the Cox model.
2013.
Springer Science & Business Media.

John D Kalbfleisch and Ross L Prentice.
The statistical analysis of failure time data, volume 360.
John Wiley & Sons, 2011.