Bucharest Academy of Economic Studies
Doctoral School of Finance and Banking

DEFAULT DETERMINANTS

in case of Romanian state-backed mortgage loans

MSc student: Elisa Maria Lung

Supervisor: PhD. Professor Moisă Altăr

June 2014

I. Overview

SCOPE

- Identifying a group of characteristics that are able to distinguish ex-ante defaulting from non-defaulting customers → case of government-insured mortgage loans
- Find whether there are / which are the macroeconomic determinants that significantly influence the probability of default of the bank's customers

METHOD

- Data source: one of the largest Romanian commercial banks
- Sample window: loans granted between Sep-2010 and Mar-2013 (19.970 accounts)
- Performance window: one year starting with the moment of loan disbursement
- Default definition: 60 days past due
- Technique: logistic regression in two approaches: frequentist and Bayesian

II. Relevant literature

Literature review

- Hand and Henley (1997) – a review of the main statistical methods used for credit scoring
- Baesens *et al* (2003) – the majority of classification techniques are quite competitive with each other
- Zandi (1998) – credit performance is influenced by the economic conditions → leading economic indicators should be included as additional variables to explain customer credit performance
- Bellotti and Crook (2007) – using survival analysis they show that the inclusion of macroeconomic variables gives a significant uplift in model's predictive performance
- Moon and Sohn (2010) – use the logistic model for predicting the SME's default; the model containing all the information available, including the economic factors, provide the most accurate scoring model

III. Initial characteristic analysis (1)

Grouping of the variables

- The socio-demographic characteristics used in the model are **grouped**

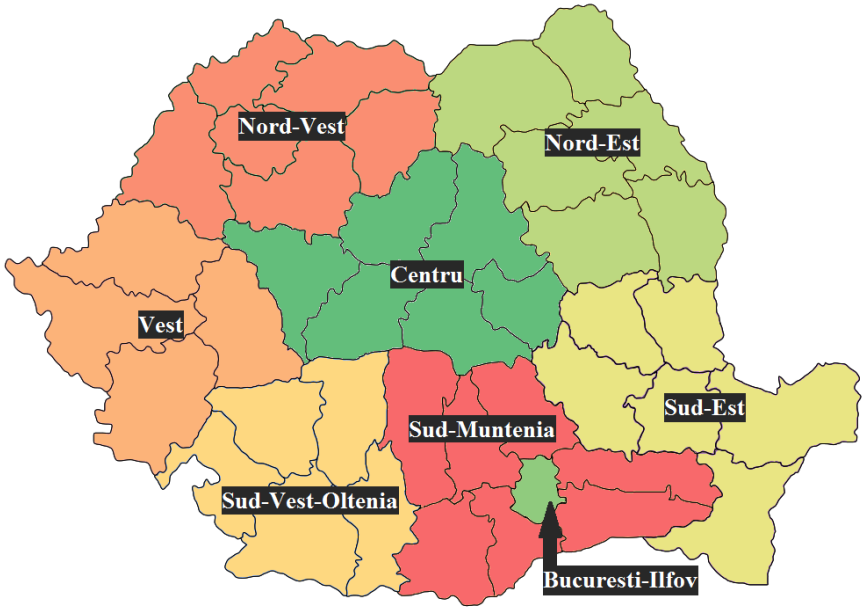
- Weight of Evidence (WOE) → measures the strength of each attribute in separating good (non-defaulted) and bad (defaulted) clients:

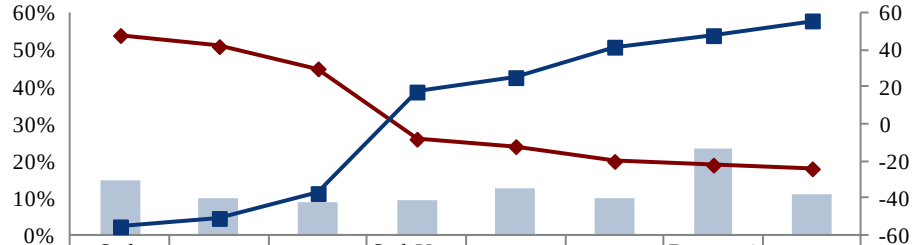
$$\left[\ln\left(\frac{\text{Distr Good}}{\text{Distr Bad}}\right)\right] \times 100$$

- WOE is used as **input** in the model, replacing the original value of the variable:
  - ✓ Outliers
  - ✓ Categorical / interval variables
  - ✓ Easier to understand relationships

III. Initial characteristic analysis (2)

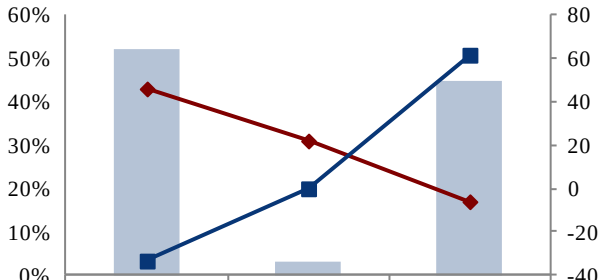
Region





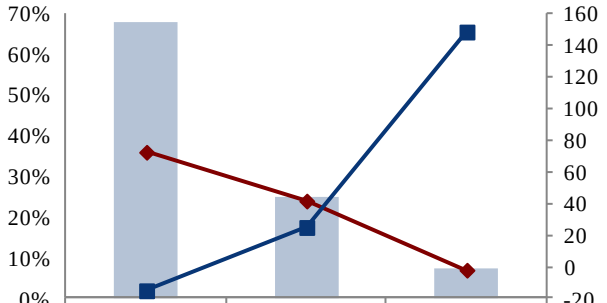
	Sud-Muntenia	Nord-Vest	Vest	Sud-Vest-Oltenia	Sud-Est	Nord-Est	Bucuresti-Ilfov	Centru
Percent	14.9%	9.7%	8.9%	9.6%	12.5%	9.8%	23.4%	11.3%
Default rate (%)	0.54	0.51	0.45	0.26	0.24	0.20	0.19	0.18
WOE	-55	-51	-37	17	25	42	48	56

Number of inhabitants



	Single person	4 or more persons	2 or 3 persons
Percent	52.1%	3.2%	44.7%
Default rate (%)	0.43	0.31	0.17
WOE	-33	0	62

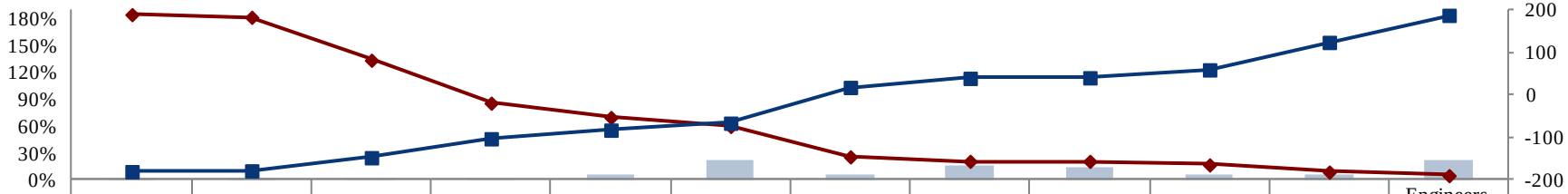
Down payment (percentage of the loan amount)



	(-;10]	(10;50]	(50;-)
Percent	68.1%	24.8%	7.1%
Default rate (%)	0.36	0.24	0.07
WOE	-15	25	149

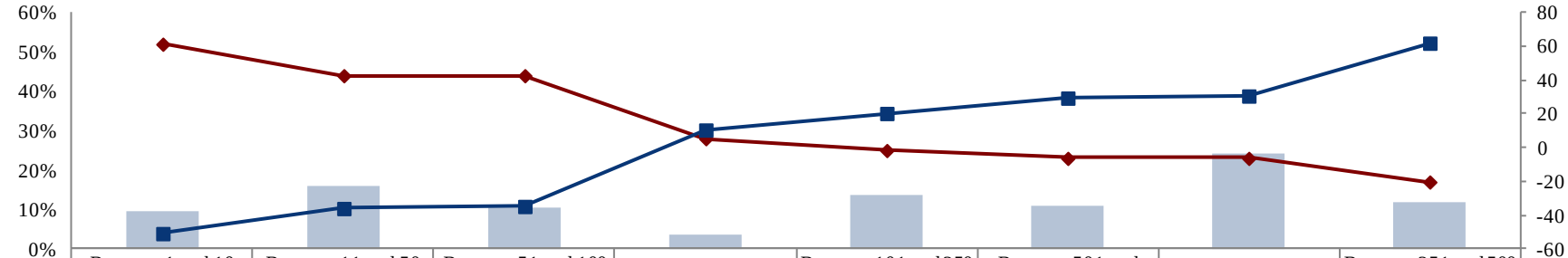
III. Initial characteristic analysis (3)

Job title



	Retired people	Freelancers	Unqualified workers	Primary-school teachers	General managers	Qualified workers	Other occupations	Public sector workers	Economists	Military personnel, police, gendarmerie	Doctors, dentists, pharmacists, veterinarians	Engineers, architects, lawyers, notaries, professors
Percent	0.3%	0.3%	1.1%	0.6%	5.7%	22.6%	5.8%	16.8%	14.5%	5.8%	5.6%	21.0%
Default rate (%)	1.85	1.82	1.34	0.86	0.7	0.6	0.26	0.21	0.21	0.17	0.09	0.05
WOE	-180	-178	-147	-103	-82	-66	18	40	41	59	124	188

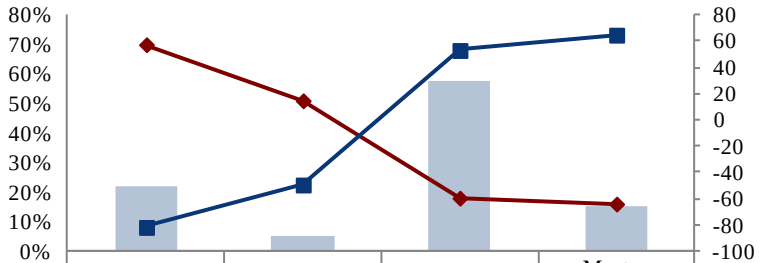
Number of employees of the borrower's current employer



	Between 1 and 10 employees	Between 11 and 50 employees	Between 51 and 100 employees	Missing	Between 101 and 250 employees	Between 501 and 1000 employees	Over 1000 employees	Between 251 and 500 employees
Percent	9.7%	15.8%	10.3%	3.6%	13.8%	10.8%	24.1%	12.0%
Default rate (%)	0.52	0.44	0.44	0.28	0.25	0.23	0.23	0.17
WOE	-51	-36	-35	10	20	29	31	62

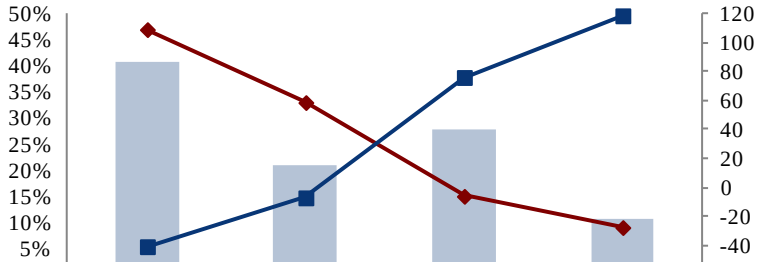
III. Initial characteristic analysis (4)

Education



	High school diploma	Vocational school	Bachelor degree	Master, professional or doctorate degree
Percent	22.2%	4.9%	57.5%	15.4%
Default rate (%)	0.7	0.51	0.18	0.16
WOE	-82	-49	53	65

Number of years spent at last / current employer



	Below 2.5 years	Between 2.5 and 4 years	Over 5 years	Between 4 and 5 years
Percent	40.8%	21.1%	27.6%	10.6%
Default rate (%)	0.47	0.33	0.15	0.09
WOE	-41	-7	76	119

Strength of a characteristic

- Weight of Evidence (WOE) – used for assessing the predictive power of each attribute
- Information Value (IV) – used for assessing the **overall** predictive power of the characteristic:

$$\sum_{i=1}^{n} (\text{Distr Good}_i - \text{Distr Bad}_i) \times \left[\ln\left(\frac{\text{Distr Good}_i}{\text{Distr Bad}_i}\right)\right]$$

Characteristic	IV	Siddiqi (2005)
Down payment	0.11	✔
Number of employees	0.14	✔
Number of inhabitants	0.20	✔
Region	0.20	✔
Years at last employer	0.28	✔
Education	0.42	✔ ✔
Job title	0.76	✔ ✔

IV. Frequentist logistic regression (1)

Model specification

- Wiginton (1980) – one of the first to use logistic regression in credit scoring

- The log of the probability odds is matched to a linear combination of the characteristic variables:

$$\log\left(\frac{p}{1-p}\right) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

- **Maximum likelihood** is used for estimating the parameters

- **Stepwise algorithm** used for variable selection → forward selection followed by backward elimination check at each step (threshold for p-value set at 0.1)

- Model validation is done using **k-fold cross validation** → development sample is divided into 10 parts

IV. Frequentist logistic regression (2)

Results

Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-5.7728	0.1547	1393.27	<.0001
WOE_REGION	1	-0.0084	0.0029	8.61	0.0033
WOE_NO_INHABITANTS	1	-0.0116	0.0032	12.91	0.0003
WOE_DOWNPAYMENT	1	-0.0085	0.0051	2.84	0.0919
WOE_EDUCATION	1	-0.0073	0.0021	11.75	0.0006
WOE_JOB_TITLE	1	-0.0077	0.0018	17.31	<.0001
WOE_YEARS_LAST_EMPLOYER	1	-0.0083	0.0028	8.77	0.0031
WOE_NO_EMPLOYEES	1	-0.0075	0.0035	4.74	0.0295

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	82.2	Somers' D	0.645
Percent Discordant	17.7	c	0.822

Comments

- We obtain a model containing seven socio-demographic characteristics
- Estimated parameters have negative signs → the result of using WOE
- An increase in WOE leads to a decrease in odds ratio (as expected)
- The statistics measuring association show the fit of the model → to be used for comparison purpose

IV. Frequentist logistic regression (3)

Selected macroeconomic variables

Unemployment rate

- An **increase** in unemployment rate prior the loan opening is expected to result in **higher** risk of default → differential is used
- Assumption: unemployment does not affect only the borrower, but also the persons related to the borrower

Output gap

- Output gap is calculated for quarterly data using Hodrick-Prescott filter
- A **positive trend** in output gap (improving economy) in the quarters preceding account opening is expected to result in **lower** default risk during the first year of loan repayment

IV. Frequentist logistic regression (4)

Results

Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-5.7248	0.1633	1229.14	<.0001
WOE_REGION	1	-0.0084	0.0029	8.54	0.0035
WOE_NO_INHABITANTS	1	-0.0117	0.0032	13.17	0.0003
WOE_DOWNPAYMENT	1	-0.0085	0.0051	2.81	0.0935
WOE_EDUCATION	1	-0.0071	0.0021	10.90	0.001
WOE_JOB_TITLE	1	-0.0076	0.0019	16.66	<.0001
WOE_YEARS_LAST_EMPLOYER	1	-0.0082	0.0028	8.56	0.0034
WOE_NO_EMPLOYEES	1	-0.0075	0.0035	4.69	0.0303
DELTA_UNEMPLOYMENT_Q1_Q3	1	1.8092	0.8677	4.35	0.0371
DELTA_GAP_Q0_Q2	1	-0.2404	0.1399	2.95	0.0856

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	82.9	Somers' D	0.659
Percent Discordant	17.1	c	0.829

Comments

- The same seven socio-demographic variables are statistically significant; similar estimated values

- Two macroeconomic variables enter the model:

✓ Δ unemployment rate – positive sign

✓ Δ output gap – negative sign

- The concordance measures are improved compared to previous version of the model

V. Bayesian logistic regression (1)

Methodology

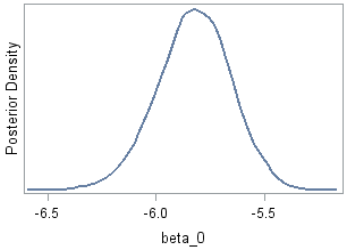
- Bayesian method offers an **alternative approach** → parameters are treated as **random** variables
- Based on simple **rules of probability** (Bayes, 1763): $p(\theta|y) = \frac{p(\theta)p(y|\theta)}{P(y)} \propto p(\theta)p(y|\theta)$
- The **posterior distribution** $p(\theta|y)$ is the density of fundamental interest
- Largest obstacle in Bayesian analysis is the integration; $p(\theta|y)$ doesn't have closed form → a **simulation method** is needed
- The alternatives are represented by **Markov Chain Monte Carlo** (MCMC) algorithms – random walk Metropolis algorithm is used

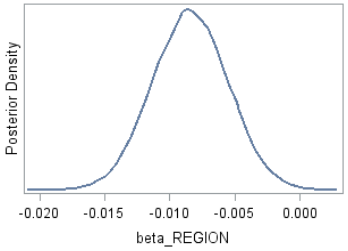
V. Bayesian logistic regression (2)

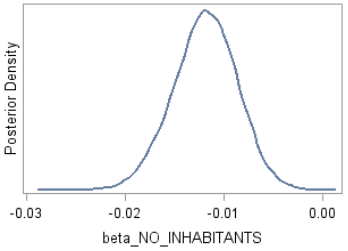
Specifications and results

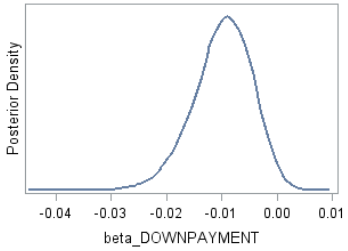
- Total number of draws: 210,000 out of which:
  - ✓ burn-in: first 10,000 observations are removed → so that stationary distribution is reached
  - ✓ thinning: only each 10th observation is kept → to reduce autocorrelation

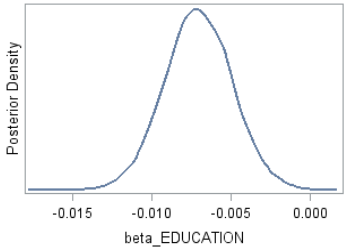
- For each parameter of interest, we have 20,000 observations drawn from posterior distribution:

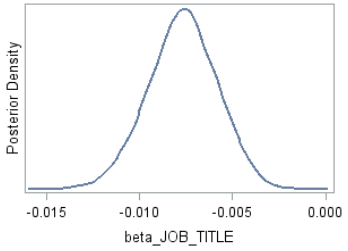


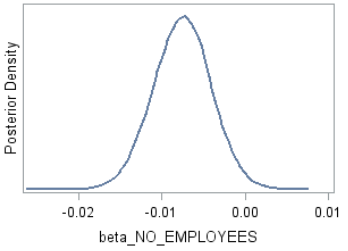


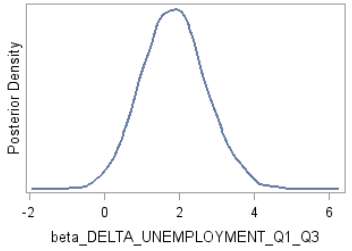


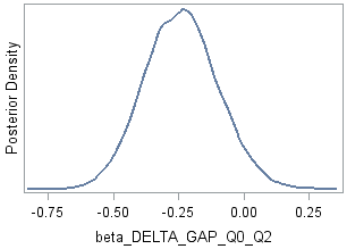












V. Bayesian logistic regression (3)

Comments

- Means of the sampled distributions are considered → very similar results when compared to classic approach
- Advantage: credibility intervals for each parameter
- Down payment → the 5% equal-tail interval does not include zero; in classic approach, p-value = 0.09
- Delta output gap → the 10% equal-tail interval does not include zero

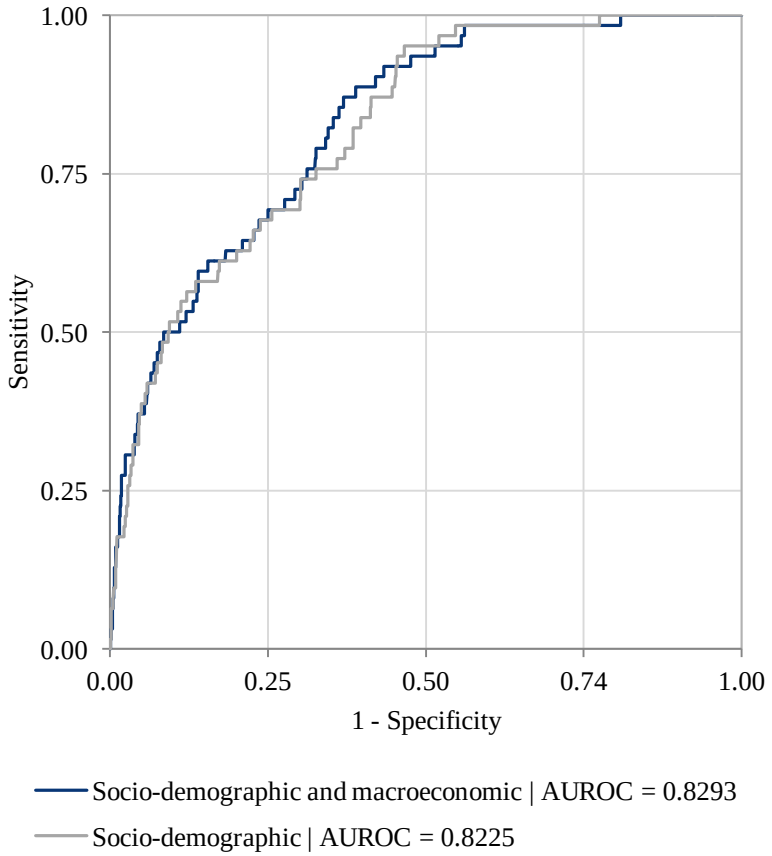
Posterior results

Parameter	Posterior Summaries			Posterior Intervals		
	N	Mean	Standard Deviation	Alpha	Equal-Tail Interval	
beta_0	20000	-5.8241	0.1690	0.05	-6.1719	-5.5033
beta_REGION	20000	-0.0084	0.0029	0.05	-0.0141	-0.0028
beta_NO_INHABITANTS	20000	-0.0119	0.0033	0.05	-0.0185	-0.0058
beta_DOWNPAYMENT	20000	-0.0098	0.0053	0.05	-0.0212	-0.0006
beta_EDUCATION	20000	-0.0071	0.0021	0.05	-0.0113	-0.0029
beta_JOB_TITLE	20000	-0.0077	0.0019	0.05	-0.0115	-0.0041
beta_YEARS_LAST_EMPLOYER	20000	-0.0085	0.0029	0.05	-0.0143	-0.0032
beta_NO_EMPLOYEES	20000	-0.0076	0.0035	0.05	-0.0144	-0.0008
beta_DELTA_UNEMPLOYMENT_Q1_Q3	20000	1.8574	0.8625	0.05	0.2138	3.6144
beta_DELTA_GAP_Q0_Q2	20000	-0.2384	0.1409	0.05	-0.5072	0.0460

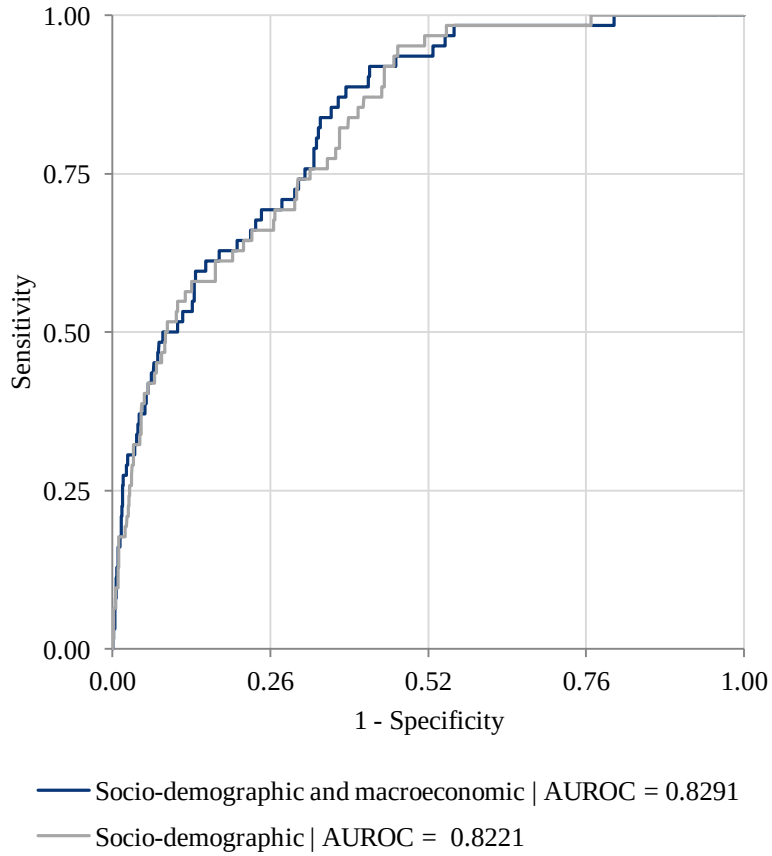
VI. Performance comparison

ROC curves for selected models

Frequentist logistic regression



Bayesian logistic regression



VII. Conclusions

Concluding remarks

- **Socio-demographic characteristics** of the borrower collected at the moment when the loan application takes place → important role in assessing the probability of default within the first year of loan repayment
- Adding **macroeconomic variables** such as evolution of unemployment rates and output gap showed a superior prediction ability of the proposed model to the initial one
- The two approaches for estimating the parameters of logistic regression (frequentist and Bayesian) provide very **similar results**; the advantage of Bayesian method lies in the fact that provides a distribution for the parameters of interest

Further research areas

- Inclusion of clients' **behavioral information** (e.g., account balance, delinquency on other loans, credit bureau information)
- Changing the objectives from trying to minimize the chance of default to **maximizing the profitability**

References (1)

[1] Avery, R. B., Calem, P. S. and Canner, G. B. (2004), "Consumer Credit Scoring: Do Situational Circumstances Matter?", *Journal of Banking and Finance*, 28(4), 835-856.

[2] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J. (2003), "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring", *Journal of the Operational Research Society*, 54(6), 627-635.

[3] Bellotti, T. and Crook, J. (2009), "Credit Scoring with Macroeconomic Variables Using Survival Analysis", *Journal of the Operational Research Society*, 60(12), 1699-1707.

[4] Campbell, T. S. and Dietrich, J. K. (1983), "The Determinants of Default on Insured Conventional Residential Mortgage Loans", *The Journal of Finance*, 38(5), 1569-1581.

[5] Chan-Lau, J. A. (2006), "Fundamentals-Based Estimation of Default Probabilities: A survey", *IMF Working Paper*, No. 149.

[6] Chandler, G. C. and Coffman, J. Y. (1983), "Applications of Performance Scoring of Accounts Receivable Management in Consumer Credit", *Journal of Retail Banking*, 5(4).

[7] Gardner, M. J. and Mills, D. L. (1989), "Evaluating the Likelihood of Default on Delinquent Loans", *Financial Management*, 18(4), 55-63.

[8] Hand, D. J. and Henley, W. E (1997), "Statistical Classification Methods in Consumer Credit Scoring: A Review", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3), 523-541.

[9] Hosmer Jr, D. W. and Lemeshow, S. (2004), *Applied Logistic Regression*, John Wiley & Sons.

[10] Jackson, J. R. and Kaserman, D. L. (1980), "Default Risk on Home Mortgage Loans: A Test of Competing Hypotheses", *The Journal of Risk and Insurance*, 47(4), 678-690.

[11] Koop, G. (2003), *Bayesian Econometrics*, John Wiley & Sons.

References (2)

[12] Liu, Y. and Schumann, M. (2005), "Data Mining Feature Selection for Credit Scoring Models", *Journal of the Operational Research Society*, 56(9), 1099-1108.

[13] Mileris, R. (2012), "Macroeconomic Determinants of Loan Portfolio Credit Risk in Banks", *Engineering Economics*, 23(5), 496-504.

[14] Moon, T. H. and Sohn, S. Y. (2010), "Technology Credit Scoring Model Considering Both SME Characteristics and Economic Conditions: The Korean Case", *Journal of the Operational Research Society*, 61(4), 666-675.

[15] Orgler, Y. E. (1970), "A Credit Scoring Model for Commercial Loans", *Journal of Money, Credit and Banking*, 2(4), 435-445.

[16] Rajaratnam, K., Beling, P. and Overstreet, G. (2010), "Scoring Decisions in the Context of Economic Uncertainty", *Journal of the Operational Research Society*, 61(3), 421-429.

[17] Rosenberg, E. and Gleit, A. (1994), "Quantitative Methods in Credit Management: A Survey", *Operations Research*, 42(4), 589-613.

[18] Siddiqi, N. (2005), *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, John Wiley & Sons.

[19] Thomas, L. C. (2000), "A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers", *International Journal of Forecasting*, 16(2), 149-172.

[20] Vandell, K. D. (1978), "Default Risk Under Alternative Mortgage Instruments", *The Journal of Finance*, 33(5), 1279-1296.

[21] Zandi, M. (1998), "Incorporating Economic Information into Credit Risk Underwriting", in Mays, E. (ed.), *Credit Risk Modeling: Design and Application*, Dearborn Publishers, Chicago/London, 155-168.

[22] Nationalbank, O. (2004), "Rating Models and Validation. OeNB Guidelines on Credit Risk Management Series", *OeNB: Vienna*.

THANK YOU

for your attention!