Credit scoring for P2P lending

Daniel Felix Ahelegbey*, Paolo Giudici

Department of Economics and Management Sciences, University of Pavia, Italy

Abstract

This paper shows how to improve the measurement of credit scoring by means of factor clustering. The improved measurement applies, in particular, to small and medium enterprises (SMEs) involved in P2P lending. The approach explore the concept of familiarity which relies on the notion that, the more familiar/similar things are, the more close they are in terms of functionality or hidden characteristics (latent factors that drive the observed data). The approach uses singular value decomposition to extract the factors underlying the observed financial performance ratios of the SMEs. We then cluster the factors using the standard k-mean algorithm. This allows us to segment the heterogeneous population into clusters with more homogeneous characteristics. The result shows that clusters with relatively fewer number of SMEs produce a more parsimonious and interpretable credit scoring model with better default predictive performance.

Keywords: Clustering, Credit Scoring, Factor Models, FinTech, P2P Lending, Segmentation

1. Introduction

When it comes to the measurement of credit scoring, the traditional concept of one model fits all may work well only for firms or individuals with capacity, credit access, cash, and/or collateral. Such models usually do not work for people with no financial history or collateral even if they have payback capabilities. Continuing with the traditional credit scoring system will not help a significant proportion of SMEs. It is therefore vital and crucial to develop alternative credit scoring models tailored in a way to allow SMEs without traditional financial history but with payback capabilities based on alternative means to have a credit score that allows them to gain access to credit.

Recent advancements gradually transforming the traditional economic and financial system is the emergence of digital-based systems. Such systems present a paradigm shift from traditional intra-organizational systems to customer-oriented technological (digital) systems. Financial technological ("FinTech") companies are gradually gaining ground in major developed economies across the world. The emergence of business-to-customer (B2C), customer-to-customer (C2C), provider-oriented business-to-business (B2B) and peer-to-peer (P2P) platforms is are typical examples of FinTech systems. Thus, Fintech's offer solutions that differ from traditional institutions regarding the providers and the interaction types as well as regarding the banking and insurance processes they support (Haddad and Hornuf, 2019; Puschmann, 2017). These platforms aims at facilitating credit services by connecting individual lenders with individual borrowers without the interference of traditional banks as

*Corresponding author

Email address: dfkahey@bu.edu (Daniel Felix Ahelegbey)

intermediaries. Such platforms serves as a digital financial market and have significantly improve the customer experience in terms of cost saving and speed of the services to both individual borrowers and lenders as well as small business owners.

Despite the various advantages current fintech systems, the existing digital platform systems inherit some of the challenges of traditional credit risk management. Credit scoring is purely based on the data available on a borrower that signals their financial worth and ability to payback loans. In addition, they are characterized by the asymmetry of information and by a strong interconnectedness among their users (see e.g. Giudici et al., 2019) that makes distinguishing healthy and risky credit applicants difficult, thus affecting credit issuers. There is, therefore, a need to explore methods that can help improve credit scoring of individual or companies that engage in P2P credit services.

This paper investigates how factor clustering-based approach to segment a population to improve the statistical-based credit score for small and medium enterprises (SMEs) involved in P2P lending. The methodology employed in this paper extends the similarity of latent factors recently adopted by Ahelegbey et al. (2019a,b). The approach explore the concept of familiarity as a signal for functional relationships among financial institutions. The familiarity concept relies on the notion that, the more familiar things are, the more close they are in terms of functionality or hidden characteristics (latent factors that drive the observed data). By this reasoning, we postulate that the more familiar the latent factors of SME performance ratios, the more close they are in terms of either holding securities with similar features, pursuing similar financial strategies or models. Such features make SMEs become more identical and increases the probability of exposure to common risk factors.

We contribute to the literature on the application of factor models in finance (see, e.g., Dungey et al., 2005; Dungey and Gajurel, 2015; Forbes and Rigobon, 2002; Fox and Dunson, 2015; Lopes and Carvalho, 2007; Nakajima and West, 2013). In this paper, we cluster the factors that drives the observed financial data, allowing us to segment the population and estimate a logistic model based on the sample segmentation.

Our empirical application contributes to modeling credit risk in SMEs particularly engaged in P2P lending. For related works on P2P lending via logistic regression, see Ahelegbey et al. (2019a,b); Andreeva et al. (2007); Barrios et al. (2014); Emekter et al. (2015); Giudici et al. (2019); Serrano-Cinca and Gutiérrez-Nieto (2016). We model the credit score of over 15000 SMEs engaged in P2P credit services across Southern Europe. We show via our empirical results that our factor clustering-based approach to segmentation presents a more efficient scheme that achieves higher performance than the conventional approach.

The paper is organized as follows. Section 2 presents the econometric methodology, Section 3 discuss application to a credit scoring database provided by a European rating agency for P2P platforms and, Section 4 presents concluding remarks.

2. Econometric methodology

2.1. Segmented Logistic Model

Suppose $Y = \{Y_i\},\ i = 1, \ldots, n$, is a vector of observations of the loan status of n firms, such that $Y_i = 1$ if firm-i has defaulted on its loan obligation, and zero otherwise. Furthermore, let $X = \{X_{ij}\},\ i = 1, \ldots, n,\ j = 1, \ldots, p$, be a matrix of n observations with p financial variables that signal the credit worthiness of institutions. The conventional approach to evaluate the conditional distribution of Y given X is a one model-fits-all logistic regression. In this application, we assume the firms can be classified into groups according to

some similarities in the latent characteristics of their observed features. Suppose there exist k non-overlapping groups of firms. We model the conditional distribution of $Y^{(l)}$ given $X^{(l)}$ for $l = 1, \ldots, k$ as a logistic model via log-odds function given by

$$\log\left(\frac{\pi_i^{(l)}}{1-\pi_i^{(l)}}\right) = \beta_0^{(l)} + X_i^{(l)}\beta^{(l)} \tag{1}$$

where $\pi_i^{(l)} = P(Y_i^{(l)} = 1|X_i^{(l)})$, β_0 is a constant term, $\beta^{(l)} = (\beta_1^{(l)}, \ldots, \beta_p^{(l)})'$ is a $p \times 1$ vector of coefficients and $X_i^{(l)}$ is the i-th row of $X^{(l)}$.

2.2. Factor Model with SVD

Let X be the observed data matrix of n institutions, each with p number of features measuring financial performance ratios. We denote with X_i, the i-th institution which corresponds to the i-th row of X. We proceed under the assumption that the observed data matrix X can be approximated via singular value decomposition given by

$$X \;=\; UDV' \tag{2}$$

where U and V are matrices of dimensions $n \times r$ and $p \times r$, $r < p$, respectively, whose columns are the left and right singular vectors of X respectively, and D is a diagonal $r \times r$ matrix which contains the square roots of the non-zero eigenvalues of $X'X$ and XX'.

2.3. Clustering Latent Coordinates

To classify the n firms into k non-overlapping groups, we consider a clustering scheme such that "similar" firms belong to the same group and "different" firms go into different groups. In this application, we use the latent coordinates of the firms in U as points in a plane (or some higher-dimensional space).

Given that U is a matrix of coordinates of n points in an r-dimensional space, the i-th row represents the coordinates of the i-th firm, while the j-th column of U represent the coordinates of the institutions on the j-th axis. For simplicity, we plot the first three dimensions of U as the default dimension of the latent positions. This agree with most applications involving multidimensional scaling and provides a convenient framework to visualize the position of agents/firms in a 3-D space.

Typical clustering methods discussed in the literature ranges from centroid-based method (k-means), to density-based, distribution-based and hierarchical methods. In this application, we follow the centroid-based method of k-means clustering.

3. Application

3.1. Data Description

To illustrate the effectiveness of the application of factor network methodology in credit scoring analysis, we obtained data from the European External Credit Assessment Institution (ECAI) on 15045 small-medium enterprises engaged in Peer-to-Peer lending on digital platforms across Southern Europe. The observation on each institution is composed of 24 financial characteristic ratios constructed from official financial information recorded in 2015. Table 1 presents a description of the financial ratios with summary of mean statistics of

Var	Formula (Description)	Active(Mean)	Defaulted(Mean)
V1	(Total Assets - Shareholders Funds)/Shareholders Funds	8.87	9.08
V2	(Longterm debt + Loans)/Shareholders Funds	1.25	1.32
V3	Total Assets/Total Liabilities	1.51	1.07
V4	Current Assets/Current Liabilities	1.6	1.06
V5	(Current Assets - Current assets: stocks)/Current Liabilities	1.24	0.79
V6	(Shareholders Funds + Non current liabilities)/Fixed Assets	8.07	5.99
V7	EBIT/Interest paid	26.39	-2.75
V8	(Profit (loss) before tax + Interest paid)/Total Assets	0.05	-0.13
V9	P/L after tax/Shareholders Funds	0.02	-0.73
V10	Operating Revenues/Total Assets	1.38	1.27
V11	Sales/Total Assets	1.34	1.25
V12	Interest Paid/(Profit before taxes + Interest Paid)	0.21	0.08
V13	EBITDA/Interest Paid	40.91	5.71
V14	EBITDA/Operating Revenues	0.08	-0.12
V15	EBITDA/Sales	0.09	-0.12
V16	Constraint EBIT	0.13	0.56
V17	Constraint PL before tax	0.16	0.61
V18	Constraint Financial PL	0.93	0.98
V19	Constraint P/L for period	0.19	0.64
V20	Trade Payables/Operating Revenues	100.3	139.30
V21	Trade Receivables/Operating Revenues	67.59	147.12
V22	Inventories/Operating Revenues	90.99	134.93
V23	Total Revenue	3557	2083
V24	Industry Classification on NACE code	4566	4624
	Total number of institutions (%)	13413 (89.15%)	1632 (10.85%)

Table 1: Description of the financial ratios with summary of mean statistics according to default status.

the institutions grouped according to their default status. In all, the data consists of 1,632 (10.85%) defaulted institutions and 13,413 (89.15%) non-defaulted companies.

Figure 1 presents a 3-D scatterplot of the SMEs latent positions based on singular value decomposition of the observed features. The coordinates of defaulted SMEs are in red circles and the non-default SMEs in green triangles.

Table 2 shows the statistics of the number and percentage of the defaulted status of SMEs based on k-means clustering of the latent coordinates. In this application, we choose $k = 2$. The table shows that 14,866 (98.81%) of the SMEs are classified in Cluster 1, while the rest 179 (1.19%) are in Cluster 2. Of those in Cluster 1, 10.80% are have defaulted and 89.20% have not. Of those in Cluster 2, 14.53% are defaulted SMEs, while 85.47% are not.

Status	Cluster 1	Cluster 2
Default	1,606 - 10.80%	26 - 14.53%
Non.Default	13,260 - 89.20%	153 - 85.47%
Total	14,866 - 98.81%	179 - 1.19%

Table 2: Statistic of defaulted status of SMEs according to k-mean clustering of the latent coordinates.

3.2. Credit Score Modeling

Table 3 reports the estimated coefficients of the logistic regression for the full sample and the clustered samples. We remark that the results of the table are derived via a thorough

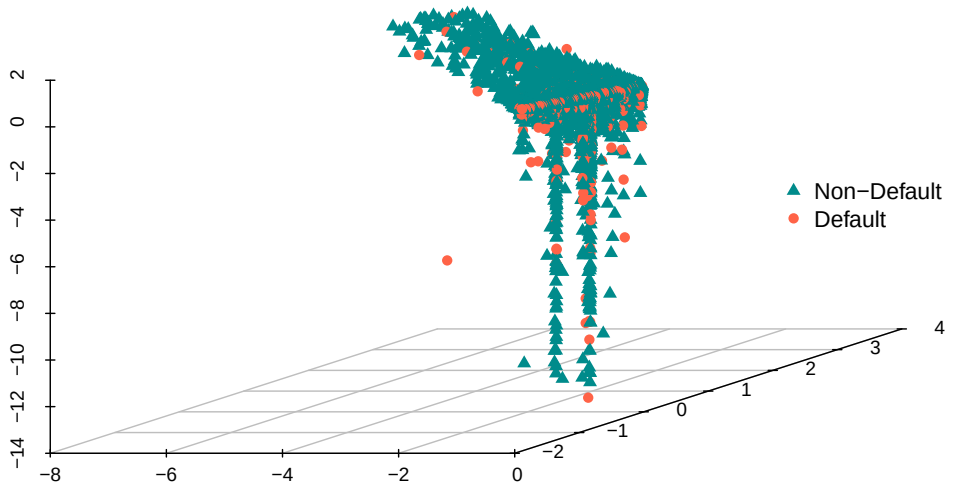

Figure 1: A 3-D scatterplot of firm's latent positions based on singular value decomposition of observed features. Coordinates of defaulted SMEs are in red circles and non-default SMEs in green triangles.

activity of model selection, aimed at obtaining the best fit statistical model using Stepwise logistic regression. The estimation of the models are carried out on the training sample which we set to be 70% of the sample. Given that 98.81% of the full sample are classified into Cluster 1, it is therefore not surprising that the credit score for the Full Sample and Cluster 1 have the same key drivers. However, we observe that for Cluster 2, the determinants of the credit score are somewhat different. There are however, some common drivers for Cluster 1 and 2, such as V3 (Total Assets/Total Liabilities), V4 (Current Assets/Current Liabilities), V14 (EBITDA/Operating Revenues), and V21 (Trade Receivables/Operating Revenues). Despite these common terms, the result shows that majority of the key drivers of credit score for those in Cluster 1 are not significant determinants for those in Cluster 2.

3.3. Comparing Default Predicting Accuracy

We evaluate the default prediction accuracy of the estimated models on the testing sample and compare the performance in terms of the standard area under the curve (AUC) derived from the receiver operator characteristic (ROC) curve. The AUC depicts the true positive rate (TPR) against the false positive rate (FPR) depending on some threshold. TPR is the number of correct positive predictions divided by the total number of positives. FPR is the ratio of false positives predictions overall negatives.

Table 4 shows the results of the area under the ROC curve of the full sample and clustered sample models. The result shows that the Full Sample model and Cluster 1 achieved 82.34% prediction rate, while Cluster 2 reported a rate of 96.77%. However, the combined performance of the Clustered Sample model attain 82.62%. Thus, the clustered sample shows a slightly higher gain in predictive performance compared to the full sample approach. This is an advantage that can be further increased considering as the cut-off the observed default percentages, which are different in the two clustered samples.

Table 4 also reports the DeLong test (DeLong et al., 1988) of the pairwise comparison of the AUC of the Full sample and that of the Clustered sample models. We perform these tests under the null-hypotheses that H_0: AUC (Full sample) $\geq$ AUC (Clustered sample) and the alternative hypotheses, H_1: AUC (Full sample) $<$ AUC (Clustered sample). The one-sided

	Full Sample	Cluster (1)	Cluster (2)
V1	0.0022	0.0023	
V2			-0.1981^{**}
V3	-0.5779^{***}	-0.5409^{***}	-4.7532^{**}
V4	-0.2761^{***}	-0.4121^{***}	-0.5076^{*}
V5		0.1882	
V6	0.0023^{*}		
V7	0.0041^{***}	0.0043^{***}	
V8	-2.2462^{***}	-2.2547^{***}	
V9			-1.3393^{**}
V10	-0.3603^{**}	-0.2687^{*}	
V11	0.4119^{***}	0.3458^{**}	2.8614
V12	0.1621^{**}	0.1610^{**}	
V13	-0.0023^{**}	-0.0024^{**}	
V14	-0.6840^{***}	-0.7531^{***}	5.0525^{***}
V15			-3.9867^{***}
V16	0.7136^{***}	0.6889^{***}	
V18	0.3775^{*}	0.4068^{**}	
V19	0.7888^{***}	0.8021^{***}	
V20		0.0007^{**}	
V21	0.0021^{***}	0.0021^{***}	0.0019^{**}
V22	0.0005^{**}	0.0005^{*}	0.0025
V23	-0.00002^{***}	-0.00003^{***}	
Constant	-2.2426^{***}	-2.3916^{***}	3.0007
Observations	12,035	11,892	143
Log Likelihood	−3,167.4570	−3,118.5110	−28.6924
Akaike Inf. Crit.	6,370.9140	6,275.0230	77.3849
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 3: Estimated coefficients from Stepwise logistic regression.

	Full Sample	Cluster (1)	Cluster (2)
AUC	0.8234	0.8234	0.9677
	Full Sample	Combine Cluster 1& 2	
AUC	0.8234	0.8262	
	Statistic	P-value	Significance
DeLong test	-1.688	0.046	**
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4: Comparing area under the ROC curve of the full sample and clustered sample models.

DeLong statistical test shows that the difference between the ROC of the Clustered sample and the Full sample is statistically significant at 95% confidence level.

In conclusion, our factor clustering approach to credit score modeling presents an efficient framework to analyze the latent positions of SMEs engaged in a P2P platform, and provides a way to segment a heterogeneous population into clusters with more homogeneous characteristics. The result shows that clusters with relatively fewer number of firms produce a more parsimonious and interpretable credit scoring model with better default predictive

performance.

4. Conclusions

This paper contributes to the strand of empirical studies to improve credit scoring for SMEs engaged in peer to peer platforms. We present a factor clustering-based approach to segment a heterogeneous population into groups with more homogeneous characteristics. The approach uses singular value decomposition to extract the factors underlying the observed financial performance ratios of the SMEs. These factors are then classified into clusters via a k-mean algorithm. We then model the credit score for each sub-population via a logistic regression.

The empirical application of our approach is demonstrated by analyzing the probability of default of over 15000 SMEs involved in P2P lending across Europe. The result shows that clusters with relatively fewer number of SMEs produce a more parsimonious and interpretable credit scoring model with better default predictive performance.

References

Ahelegbey, D. F., P. Giudici, and B. Hadji-Misheva (2019a). Factorial Network Models To Improve P2P Credit Risk Management. *Frontiers in Artificial Intelligence 2*, 8.

Ahelegbey, D. F., P. Giudici, and B. Hadji-Misheva (2019b). Latent Factor Models For Credit Scoring in P2P Systems. *Physica A: Statistical Mechanics and its Applications 522*, 112–121.

Andreeva, G., J. Ansell, and J. Crook (2007). Modelling Profitability Using Survival Combination Scores. *European Journal of Operational Research 183*(3), 1537–1549.

Barrios, L. J. S., G. Andreeva, and J. Ansell (2014). Monetary and Relative Scorecards to Assess Profits in Consumer Revolving Credit. *Journal of the Operational Research Society 65*(3), 443–453.

DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics 44*(3), 837–845.

Dungey, M., R. Fry, B. González-Hermosillo, and V. L. Martin (2005). Empirical Modelling of Contagion: A Review of Methodologies. *Quantitative Finance 5*(1), 9–24.

Dungey, M. and D. Gajurel (2015). Contagion and Banking Crisis–International Evidence for 2007–2009. *Journal of Banking and Finance 60*, 271–283.

Emekter, R., Y. Tu, B. Jirasakuldech, and M. Lu (2015). Evaluating Credit Risk and Loan Performance in Online Peer-to-Peer (P2P) Lending. *Applied Economics 47*(1), 54–70.

Forbes, K. J. and R. Rigobon (2002). No Contagion, Only Interdependence: Measuring Stock Market Comovements. *The Journal of Finance 57*(5), 2223–2261.

Fox, E. B. and D. B. Dunson (2015). Bayesian Nonparametric Covariance Regression. *The Journal of Machine Learning Research 16*(1), 2501–2542.

Giudici, P., B. Hadji-Misheva, and A. Spelta (2019). Network Based Scoring Models to Improve Credit Risk Management in Peer to Peer Lending Platforms. *Frontiers in Artificial Intelligence 2*, 3.

Haddad, C. and L. Hornuf (2019). The Emergence of the Global Fintech Market: Economic and Technological Determinants. *Small business economics 53*(1), 81–105.

Lopes, H. F. and C. M. Carvalho (2007). Factor Stochastic Volatility with Time Varying Loadings and Markov Switching Regimes. *Journal of Statistical Planning and Inference 137*(10), 3082–3091.

Nakajima, J. and M. West (2013). Bayesian Analysis of Latent Threshold Dynamic Models. *Journal of Business and Economic Statistics 31*(2), 151–164.

Puschmann, T. (2017). Fintech. *Business & Information Systems Engineering 59*(1), 69–76.

Serrano-Cinca, C. and B. Gutiérrez-Nieto (2016). The Use of Profit Scoring as an Alternative to Credit Scoring Systems in Peer-to-Peer Lending. *Decision Support Systems 89*, 113–122.