Best WI-FI 2024 - FinTech Contest

Statistical Analysis of FinTech Data: Insights from Unsupervised to Supervised Approaches

Antonio Cola & Rosario Urso

University of Naples Federico II

28 June 2024

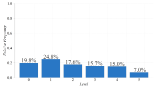
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

FinTech Dataset and Metodologies

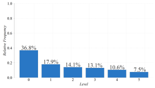
- The dataset consists of 78 variables of different types and 625 observations, each representing a data unit collected for each variable;
- The aim of this work is to identify the characteristics that determine the level of knowledge and use of financial technology tools through a dual approach:
  - Unsupervised Approach: iteractive Factor Clustering of Binary Data;
  - Supervised Approach: Graded Response Model (IRT) and Regularized Multiple Linear Regression.

Item Distribution

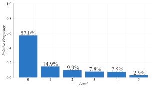
The level of knowledge and use of financial technology tools was measured through 32 items divided into four macro categories: Digital Services, Modern Tools, Investment Opportunities and General Financial Knowledge.



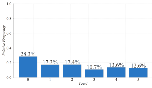
AI



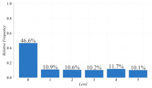
Biometric



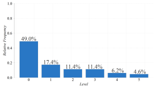
Blockchain



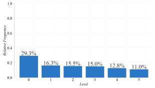
Cloud Computing



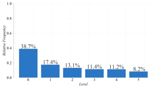
Crowdfunding



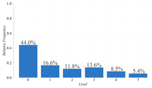
Cryptocurrencies



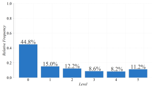
Instant Payments



IoT



Machine Learning



P2P

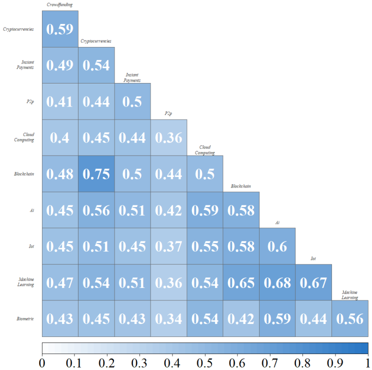
Variable Distribution

Understanding the distribution of variables in a dataset is important because it enables us to grasp the data's structure and assess the presence of anomalies or patterns.
The distributions examined concern the most interesting covariates:

Variable	Categories	n	%
Gender			
	Female	346	55.36
	Male	279	44.64
Components			
	Over 3	414	66.24
	Up to 3	211	33.76
Livewith			
	Not Alone	34	05.44
	Alone	591	94.56
Region			
	Center-North	308	49.28
	South	317	50.72
Education			
	Graduated	312	49.92
	Not Graduated	313	50.08
Area			
	STEM	310	49.60
	Not STEM	315	50.40
FedericoII			
	Not Unina	422	67.52
	Unina	203	32.48
Sector			
	Not STEM	509	81.44
	STEM	116	18.56

Correlation and Cronbach's Alpha

The correlation values of the items are all positive, ranging from a minimum of 0.34 to a maximum of 0.75.



Cronbach's Alpha Coefficient measures how different questions or items are correlated with each other, providing an indication of internal cohesion.

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_{y_i}^2}{\sigma_y^2}\right) = 0.91$$

However, a better Cronbach's alpha value can be obtained by excluding certain items.

iteractive Factor Clustering of Binary Data (i-FCB)

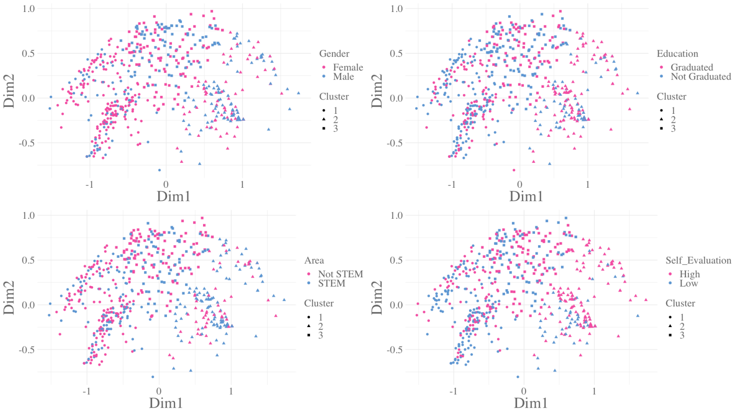
- Iterative Factor Clustering of Binary Data (Iodice D'Enza et al., 2013) is an algorithm designed to handle the clustering of binary data.
- It integrates Non-Symmetric Correspondence Analysis (NSCA) with K-Means clustering in order to improve the clustering performance.
- The function to minimize is:

$$\underset{\mathbf{B},\mathbf{Z}_H,\mathbf{G}}{min} = \left|\left|\mathbf{Z}_H'\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-1} - \mathbf{G}\mathbf{B}'\right|\right|^2 + ||\sqrt{nq}\mathbf{D}_w\mathbf{M}\mathbf{Z}\mathbf{B} - \mathbf{Z}_H\mathbf{G}||^2$$

under the orthonormality constraint $\mathbf{B}'\mathbf{D}_z\mathbf{B} = nq\mathbf{I}_h$.

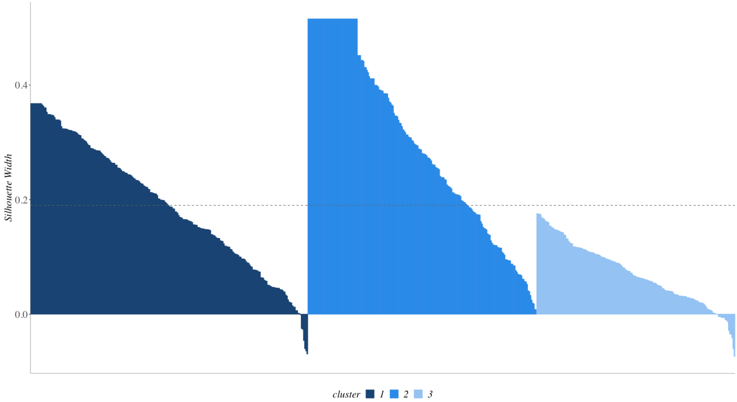
Cluster Distribution

The algorithm identified the presence of 3 clusters, characterized by different socio-demographic variables:



Silhouette Analysis

In order to check the optimal number of clusters and whether the groups were well separated, silhouette analysis was used:



Item Selection

- In the item selection process, the 'backward' method was employed.
- This approach led to the identification of a combination of items as the best for measuring the latent construct: 'Knowledge of FinTech Tools'.
- This combination includes the following items: cryptocurrencies, blockchain, AI, IoT, and machine learning, selecting a total of 5 items out of the 10 initially considered.

Graded Response Model (IRT)

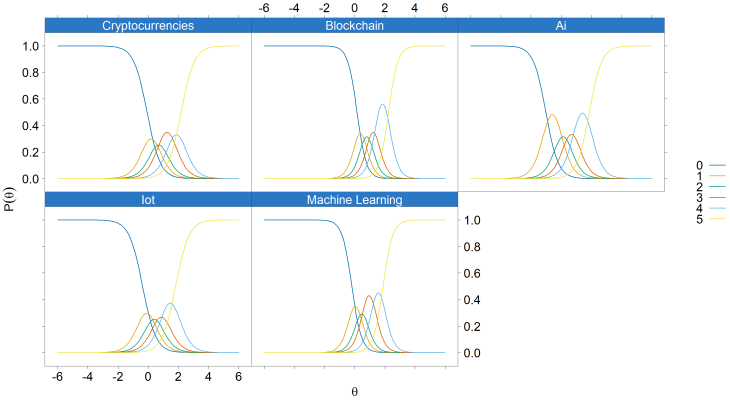
The Graded Response Model (Samejima, 1969) is defined as follows:

$$P(Y_{pi} \geq r | \theta_p, \alpha_i, \delta_i) = F(\alpha_i(\theta_p - \delta_{ir})), \quad r = 1, \ldots, k$$

- θ_p are defined as person parameters and indicate the ability level of each individual;
- δ_{ir} are defined as location parameters and indicate the position of each item on the continuous latent trait;
- α_i are defined as discrimination parameters and indicate how well an item can discriminate between different ability levels.

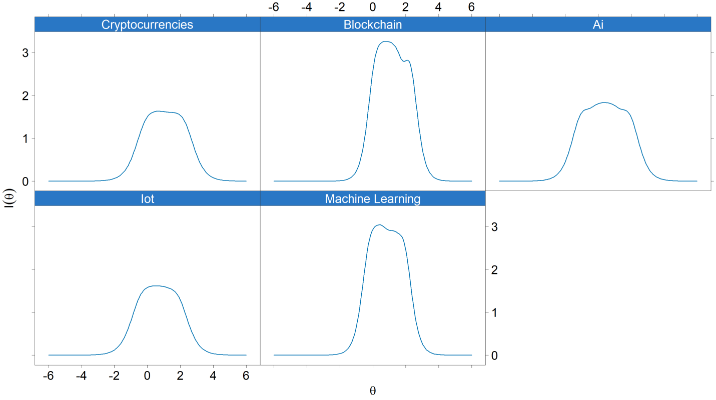
Categories Response Curves

It is interesting to examine the probabilities of responding to specific categories in an item's response scale. These probabilities are graphically displayed in the category response curves (CRCs).



Item Information Function

In polytomous models, the amount of information an item contributes depends on its slope parameter: the larger the parameter, the more information the item provides.



Regularized Multiple Linear Regression

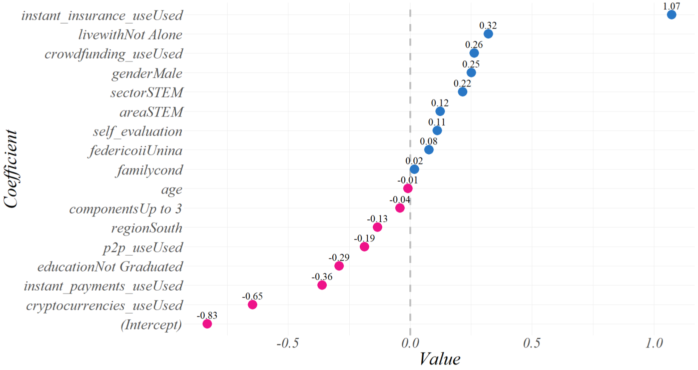
The first step involved adopting shrinkage methods for the selection of explanatory variables. In particular LASSO, Ridge and Elastic Net Regression.

$$\min_{\beta_0,\beta} = \left(\sum_{i=1}^{n} (y_i - \beta_0 - \beta x_i^T)^2 + \lambda \left(\gamma \sum_{j=1}^{m} |\beta_j| + \frac{(1-\gamma)}{2} \sum_{j=1}^{m} \beta_j^2 \right) \right)$$

Method	γ	λ	**AIC**	**BIC**
LASSO	1	0 .0085	-129 .3925	-58 .3885
Ridge	0	0 .1181	-123 .9991	-44 .1196
Elastic Net	0 .94	0 .0045	-129 .8782	-58 .8742

Elastic Net Coefficients

The objective is to identify the linear relationship between the variables, so that it can be utilized to make predictions regarding the dependent variable based on the values of the explanatory variables.



Conclusion

- The unsupervised approach identified three clusters based on FinTech knowledge, characterized by socio-demographic variables favoring males, graduates, STEM students, and high self-assessment.
- The supervised approach confirmed these results and added factors like younger age and Central-Northern residency.
- Both approaches consistently highlight the impact of various variables on FinTech knowledge.

References

[1] A. Agresti. *Analysis of Ordinal Categorical Data.* Wiley, 2 edition, 2010.

[2] A. Agresti. *Categorical Data Analysis.* Wiley, 3 edition, 2013.

[3] F. B. Baker, S. Kim, and Others. *The Basics of Item Response Theory Using R.* Springer, 2017.

[4] R. P. Chalmers. mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, 48, 2012.

[5] A. Cola, M. Iannario, and G. Tutz. Item selection method in irt models. Technical article, 2024.

[6] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[7] A. Dobson. *An Introduction to Generalized Linear Model.* Springer, 2 edition, 1990.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[9] T. Hastie and H. Zou. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

[10] A. E. Hoerl and R.W. Kennard. Ridge regresssion: Biased estimation for nonorthogonal problems. *Technometrics,*, 12:55–67, 1970.

[11] M. Iannario, A. C. Monti, D. Piccolo, and E. Ronchetti. Robust inference for ordinal response models. *Electronic Journal of Statistics*, 11(2):3407–3445, 2017.

[12] A. Iodice D'Enza, A. Markos, and M. van de Velden. Beyond tandem analysis: Joint dimension reduction and clustering in r. *Journal of Statistical Software*, 91:1–24, 2019.

[13] A. Iodice D' Enza and F. Palumbo. Iterative factor clustering of binary data. *Computational Statistics*, 28:789–807, 2013.

[14] A. Iodice D' Enza, F. Palumbo, and M. Van de Velden. Cluster correspondence analysis. *Psychometrika*, 82:158–185, 2017.

[15] G. James, D. Witten, T. Hastie, R. Tibshirani, and Others. *An introduction to statistical learning*, volume 112. Springer, 2013.

[16] C. Li and M. Hansen. Limited information goodness of fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, 66(2), 2013.

[17] G. N. Masters and B. D. Wright. The essential process in a family of measurement models. *Psychometrika*, 49:529–544, 1984.

[18] K. P. Murphy. *Machine Learning: A Probabilistic Perspective.* MIT press, 1 edition, 2012.

[19] L. Palazzo, M. Iannario, and F. Palumbo. Integrated assessment of financial knowledge through a latent profile analysis. *Behaviormetrika*, 51:319–339, 2024.

[20] D. Piccolo. *Statistica.* Il Mulino, 3 edition, 2010.

[21] F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, 1969.

[22] F. Samejima. Graded response models. In *Handbook of item response theory*, pages 95–107. Chapman and Hall/CRC, 2016.

[23] G. Schauberger. *MultOrdRS: Model Multivariate Ordinal Responses Including Response Styles*, 2024. R package version 0.1-3.

[24] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[25] W. G. Smith. Does gender influence online survey participation? a record-linkage analysis of university faculty online survey response behavior. *Online submission*, 2008.

[26] D. Thissen and L. Steinberg. A taxonomy of item response models. *Psychometrika*, 51(4):567–577, 1986.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–88, 2010.

[28] G. Tutz. *Regression for Categorical Data.* Cambridge University Press, 2012.

[29] G. Tutz. On the structure of ordered latent trait models. *Journal of Mathematical Psychology*, 96, 2020.

[30] G. Tutz. Ordinal regression: a review and a taxonomy of models. *WIRES Computational Statistics, to appear*, 2021.

[31] G. Tutz. Ordinal regression: A review and a taxonomy of models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(2):e1545, 2022.

Thanks!