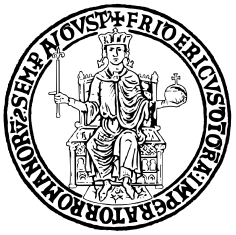
University of Naples Federico II



Political Sciences Department

Master's Degree Program in
Statistical Sciences for Decision Making

Best WI-FI 2024 Young Researcher Paper – FinTech Contest

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

Supervisor:
Professor Alfonso Iodice D'Enza

Researchers:
Antonio Cola & Rosario Urso

Academic Year 2023/2024

Abstract

This study aims to examine and analyze data from a survey aimed at identifying gender differences in the understanding and usage of financial technologies (FinTech). The first phase involved an exploratory data analysis, which examined the distributions of the variables of interest, assessed the correlations among the items (dependent variables), and calculated Cronbach's alpha to assess the internal consistency of the items. Subsequently, the work was divided into two distinct approaches: unsupervised and supervised. In the unsupervised approach, the Iterative Factor Clustering of Binary Data algorithm was used, involving alternating phases of dimensionality reduction, via Non-Symmetric Correspondence Analysis, and clustering, via K-Means. The supervised approach was preceded by a phase of item selection using the Backward method (Cola et al., 2024 [5]). The selected items, defining the latent construct "Knowledge of FinTech Tools", were used to estimate a Graded Response Model of Item Response Theory. Subsequently, the latent variable theta, representing the latent construct, was extracted from the model and used as the response for multiple regression. The covariates included in the model were selected using Shrinkage methods such as LASSO, Ridge, and Elastic Net. The results obtained from both methods (section: 5 Conclusion) are consistent and confirm the expected trend in the distribution of the phenomenon "Knowledge of FinTech Tools."

Keywords: Backward Method, Clustering, Correlations, Cronbach's Alpha, Dimensionality Reduction, Elastic Net, Graded Response Model, Item Response Theory, Iterative Factor Clustering of Binary Data, K-Means, Knowledge of FinTech Tools, LASSO, Latent Variable, Multiple Regression, Non-Symmetric Correspondence Analysis, Ridge, Shrinkage Methods, Supervised Approach, Unsupervised Approach.

1 Introduction

The statistical survey under consideration aims to assess the level of knowledge and usage of FinTech tools by the selected sample. The results of this research will provide a precise picture of the spread and usage of these innovative financial technologies and will help to understand how people interact with them. Data collection was carried out through an online questionnaire. This survey is of particular importance in an increasingly digital and rapidly evolving world, where FinTech is playing an increasingly significant role in the financial industry.
The dataset consists of 78 variables and 625 observations.

2 Exploratory Analysis

This study aims to analyze a wide range of variables relevant to the contemporary financial landscape, focusing primarily on the evolution and impact of financial technologies (FinTech). The considered variables encompass various forms of financial innovation, such as crowdfunding, cryptocurrencies, instant payments, automated advisory services (roboadvisors), as well as key concepts like artificial intelligence (AI), the Internet of Things (IoT), and machine learning. Additionally, traditional investment-related variables are examined, including bonds, stocks, and mutual funds, along with crucial themes such as monetary policy, risk, and return. To contextualize the analysis, demographic and socio-economic factors are also taken into account, such as age, gender, education level, occupation, income, and the geographical location of participants. The main objective is to provide a comprehensive overview of contemporary financial dynamics, examining both traditional

elements and emerging trends within the context of FinTech.
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. This, in turn, can aid in making decisions regarding the analysis techniques to employ and formulating reliable conclusions about the dataset.
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

Table 1: Variable Distribution

3 Unsupervised Approach

3.1 Iterative Factor Clustering of Binary Data (i-FCB)

In the framework of unsupervised methods, it is sometimes useful to identify an approach that simultaneously reduces dimensionality and characterizes one or more groups of statistical units. Iterative Factor Clustering of Binary Data (i-FCB) algorithm consists of integrating Non-Symmetric Correspondence Analysis (NSCA) to K-Means clustering (Iodice D'Enza et al., 2013 [13], 2017 [14], 2019 [12]).
Denoting with $\mathbf{Z_H}$ a matrix with memberships of objects in H clusters, $\mathbf{B_j}$ a matrix of weights and G a cluster centroid matrix, the function to minimize is shown below:

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

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

This method then performs a CA where the dependent variable is the cluster membership indicator $\mathbf{Z_H}$ and the explanatory variables are the p categorical variables $\mathbf{Z_j}$. Then, after centering and weighting the $\mathbf{Z_j}$-matrix containing the dummy-coded variables, K-Means clustering is applied to the subjects' scores in order to obtain the largest possible variance between the cluster means.
This method appears to be particularly efficient when the numerosity within the groups is balanced and provides reliable results by successfully qualifying the clusters.
To identify the optimal number of clusters and check whether they were well separated, Silhouette analysis was performed. With such a graph, the distance of one point from all others is calculated, and the fact that all values are positive means that all units have been well classified.
The obtained clusters were characterized by a set of socio-demographic variables after testing the association with a χ^2 test with the classification variable:

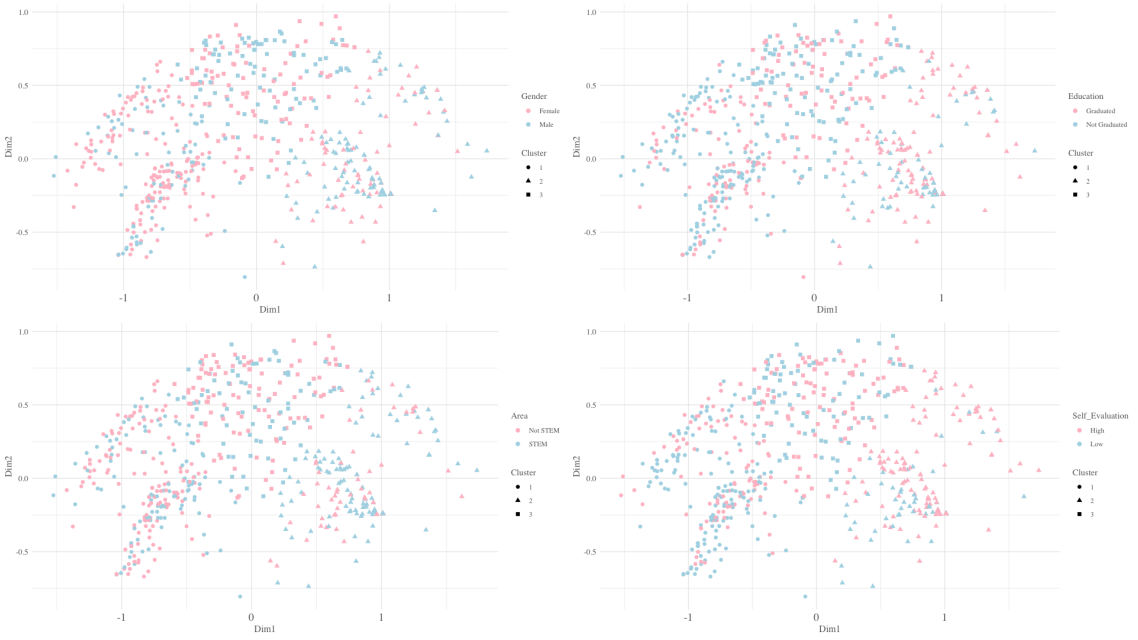


Figure 1: Clusters from i-FCB algorithm

4 Supervised Approach

4.1 Graded Response Model (Item Response Theory)

In numerous applications, items are characterized by more than two response categories, which are arranged in a specific order. Consider items scored polytomously, represented by values Y_{pi} which belong to the set $\{0,1,\ldots,k\}$, where p ranges from 1 to P, and i from 1 to I. The set $\{0,1,\ldots,k\}$ symbolizes the sequential arrangement of these categories. For the sake of clarity, we adopt a constant number of response categories, $k+1$, when formulating the models. However, it should be noted that the actual number of response categories may differ among items, with the specific categories for item i being denoted as $\{0,1,\ldots,k_i\}$.

One of the most widely used ordinal models is Samejima's Graded Response Model (Samejima, 1995 [21], 2016 [22]). This model consists of independent binary models for categorizing responses. Imagine the response categories denote levels of achievement in a

test. These can be divided into two groups: $\{0,1,\ldots,r-1\}$ representing low performance and $\{r,\ldots,k\}$ for high performance, with "low" and "high" indicating performances "below category r" and "at or above category r", respectively. The division into low and high performance groups is defined using a binary model that incorporates the individual's ability, θ_p (in this case: "Knowledge of FinTech Tools"), and a threshold that varies depending on the category where the split occurs.

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

Therefore, for each split variable, the dichotomization into categories $\{0,1,\ldots,r-1\}$ and $\{r,\ldots,k\}$, a binary model is presumed to apply. Crucially, these models operate concurrently, utilizing the same individual ability θ_p but with differing location parameters δ_{ir}. A straightforward reformation leads to the cumulative model.

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

In the formulation, the location parameters are assembled into the vector $\delta_i^T = (\delta_{i1},\ldots,\delta_{ik})$. Originally introduced by Samejima as a normal-ogive model, the function $F(.)$ represented the normal distribution. The logistic adaptation employs the logistic function $F(.)$.
The model described is built from binary models based on the dichotomies formed by the categories $Y_{pi} < r$ and $Y_{pi} \geq r$. As a result, the location parameters are ordered, necessitating that $\delta_{ir} \leq \delta_{i,r+1}$ for all categories.

$$P(Y_{pi} = r) = P(Y_{pi} \geq r) - P(Y_{pi} \geq r+1) = F(\alpha_i(\theta_p - \delta_{ir})) - F(\alpha_i(\theta_p - \delta_{i,r+1})) \geq 0$$

Null Hypothesis	Alternative Hypothesis	Critical Region
H_0 : X and Y are Indipendent	H_0 : X and Y are not Indipendent	$RC(\alpha) : X^2 > \chi^2_{(\alpha,g=(k-1)(h-1))}$

La statistica del test, anche conosciuta come Test del Chi-Quadrato, può essere calcolata come segue:

	M2	df	p	RMSEA	RMSEA_5	RMSEA_95	SRMSR	TLI	CFI
Stats	3.20	5	0.67	0	0	0.03	0.02	1.20	1

4.2 Multiple Linear Regression

The Linear Regression Model is a statistical method used to examine the relationship between a dependent variable (or response) and one or more independent variables (or predictors). The goal is to identify and quantify the linear relationship between the variables so that it can be used to make predictions about the dependent variable based on the values of the independent variables. The regression coefficients provide information about the strengths and directions of these relationships.
The first step involved adopting shrinkage methods for the selection of explanatory variables. In particular LASSO, Ridge and Elastic Net regression.

LASSO (Least Absolute Shrinkage and Selection Operator (Tibshirani, 1996 [27])) is a regularization approach used in statistical regression, useful for creating parsimonious models through variable selection. It is used to minimize the sum of squared residuals (RSS) in linear regression and for maximizing the penalized log-likelihood function in

Maximum Likelihood Estimation (MLE), which is beneficial in high-dimensional data scenarios. The LASSO regression, described as:

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

where λ is the regularization parameter, encourages models with many zero coefficients, aiding in variable selection. However, the selection of λ is critical, often determined through cross-validation, and LASSO may ignore multiple correlated variables.

Ridge regression (A. E. Hoerl et al., 1970[10]), also known as Tikhonov regularization, combats multicollinearity in statistical models by adding a penalty proportional to the square of the coefficients. In RSS, Ridge aims to minimize the residual sum of square adding a penalty term:

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

This method reduces coefficient magnitude, helping to address multicollinearity issues and produce more stable estimates, particularly when the number of variables is large relative to observations. Unlike LASSO, Ridge regression does not lead to variable selection but is favored when all variables are relevant and multicollinearity needs to be managed.

An alternative method, named Elastic Net (J. Friedman et al., 2010 [8], T. Hastie et al., 2005 [9]), is obtained combining LASSO and Ridge regression as follows:

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

where γ is a fundamental term that balances the relative importance of the two penalization techniques used: LASSO and Ridge Regression and λ is obtained through Cross-Validation approach.

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

Table 2: Method Comparison

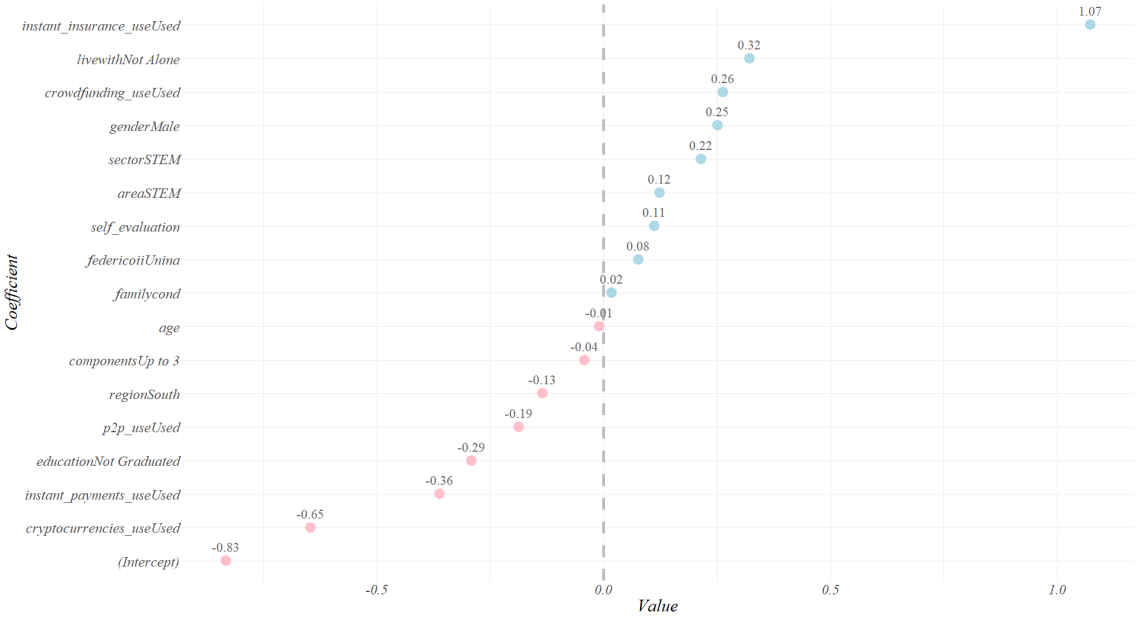


Figure 2: Elastic Net Coefficients

5 Conclusion

In conclusion, the unsupervised approach revealed three clusters: Cluster 1, comprising individuals with low or absent knowledge of FinTech tools; Cluster 3, consisting of individuals whose knowledge is limited to more common or simpler tools; and Cluster 2, including individuals with more extensive or comprehensive knowledge of FinTech tools. These groups were then characterized by socio-demographic variables, with the emerged results indicating a difference in FinTech knowledge favoring males, graduates, STEM students, and those with a high self-assessment level.
On the other hand, the supervised approach highlighted the following results: a difference in FinTech knowledge favoring males, graduates, STEM workers, STEM students, high self-assessment level, living with others (not alone), younger age, and residents of the Central-Northern regions.
By comparing the results obtained from both approaches, a consistency is observed, underscoring the accuracy of the findings and allowing for an understanding of how the variability of the "knowledge of FinTech tools" phenomenon changes with the impact of different variables.

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.