Course of Data Science Lab - Professor M. Fattore and S. Gerli - Academic year 2023/2024

Stereotypes and fraud: can common knowledge identify fragile subjects?

Carlo Arpini 918543 - Emmanuele Lotano 918608 - Chiara Mariani 918354

Università degli Studi di Milano-Bicocca, CdLM Data Science

Abstract

Within the digital age, financial frauds have found new fertile terrain thanks to newly birthed products and ease of use of instruments that move capital. Our research tries to identify links between common financial attributes and one's fragility and vulnerability to financial frauds through a 2017 survey proposed by Bank of Italy to which we applied Machine Learning methods. The results highlight some common factors that are within the domain of public knowledge but also a systematic aspect to frauds that applies to everyone and other aspects that encompass the Italian landscape, such as a low financial knowledge and a debt averse society.

Contents

1 Introduction

1.1 Context of our analysis

In the digital context in which Italian citizens live every day, information technology has transformed each of our lives into an existence rooted in an online environment. This has brought with it not only benefits, such as the speed of information propagation and the possibility of immediate global comparison, but also numerous harms, including a pathological dependence on digital and an impoverishment of human relationships. A further negative aspect of digital society that has been on the rise in recent years is the issue of online financial fraud. The advent of online transactions, the emergence of e-banking and the digitalisation of international economies have increased the risk of being a victim of online fraud, affecting a significant portion of the population, regardless of age, gender or social background.

In Italy, cyber frauds are considered a phenomenon that constantly threatens customers and banks. In this regard, Corriere del Mezzogiorno [1] specified that, based on data from the CRIF Observatory Report, there has been an increase in both the average amount of frauds and a change in criminal strategies. In particular, the regions most affected were Lombardy with 13.8% of total assets stolen with frauds, Sicily with 13.2%, Campania with 12.7%, Lazio and Puglia trailing close just under 10% with 9.3% and 8.9% respectively. These high percentages highlight not only the ability of scammers to manipulate victims, but also the vulnerability of a good portion of the Italian population to such threats.

CONSOB [2] emphasizes that online frauds pose a threat to financial stability and consumer confidence, as savings are a valuable asset that should not be evaporated in the hands of scammers. Therefore, it is necessary to take preventive measures, such as promoting greater awareness and financial education of citizens.

Using a questionnaire designed by the Bank of Italy [3] in 2017 and completed by approximately 2,500 individuals, we attempt to discern which factors determine whether or not an individual can be considered susceptible to exploitation. It will be of interest to evaluate whether the findings align with the collective imagination or deviate from conventional wisdom due to cognitive biases to guide any regulatory standpoint.

1.2 Idea and aim of our analysis

In the context of an environment where online frauds are a persistent and growing concern, it is crucial to identify any pattern that might be linked to being victims of a financial fraud, to act upon them. This is the central question that we address in our analysis.

The aforementioned questionnaire includes questions of a socio-demographic nature, such as geographic area of origin, educational qualification or employment status. It also includes questions related to financial knowledge, such as the definition of simple and compound interest or inflation. Additionally, it encompasses questions related to each individual's financial behavior, touching on topics such as savings, long-term goals or the purchase of financial instruments.

As a preliminary step, we identified the question in the questionnaire that pertains to whether an individual has ever been the victim of a fraud. Once the preliminary steps had been completed, enabling the dataset to be processed in a suitable manner, an attempt was made to consolidate the various questions into eight subcategories, thus facilitating interpretation. Three different algorithms (Random Forest, Logistic Regression and Multi Layer Perceptron) were applied to try and classify fraud victims based on the Matthews Correlation Coefficient [4], [5] on both the full dataset and the dataset implemented with feature engineering. Subsequently,

undersampling techniques and class weights were employed to address some inherent problems in the data distribution. As there was a small but acceptable reduction in predictive capacity, the dataset obtained through feature engineering was selected for the final analysis given it had sufficient results but great explainability. The most significant features, as determined by Gini entropy decrease and Permutation Importance, and the relationship between each and the target variable can be investigated through Partial Dependences of each feature. These methods enabled us to forge a path in interpreting stereotypical ideas and values in the personal financial space, but ultimately demonstrated the capillary presence of frauds for all classes of victims.

Ultimately, we also sought to construct six prototypical profiles of individuals who might have responded to our survey based on common knowledge and common ideas to assess the likelihood of them being defrauded and to gain insights on if the stereotypes hold true or not under our analysis and what can be gained from a regulatory "prevention - based" approach standpoint.

2 Dataset Description

The dataset is essentially comprised of 2376 replies by 106 columns, which can be seen in the Appendix 10.1. We have such a high number of columns as multiple choice questions are represented by binary columns, which have values of ones when the choice is selected and zeros otherwise; this results in a number of columns that is almost three times that of the questions highlighted in the Appendix 10.1. Because of this, any analysis done on the dataset as is is essentially uninterpretable, and we are forced to first try and clean/reduce our dataset to enhance interpretability of even the most basic analysis.

Moreover, we do not have a direct target variable but rather we have three questions that can be used to infer values of the target variable; this also calls for more data handling.

2.1 Main pain points

There are also some other pain points to be highlighted. First things first, not all questions were asked to participants. As an example, question *"qd12"* triggers only whenever the participant replies a specific reply to question *"qd11"*. As a result we have also to deal with a lot of NaN values, with the upside that they are easily interpretable; note that this is not the only example of this phenomenon in the dataset.

At last, there are also pain points which will be present until the end by definition and will not be solvable: as an example, there is no interpretation for the *"pesofitc"* column, nor we have any way of telling if a participant is truthful or not in its replies, and we may have inconsistency, which will be almost impossible to detect unless in reading and reinterpreting all the data points. Some information from the original research as such will either not be accessible and some level of uncertainty will systematically be present in the analysis.

3 Data cleaning

The data cleaning section will focus on handling of NaN values and a light feature reduction that will not result in losing any information that is to be considered relevant. This is after, most importantly, we recreate our target variable: a binary class called *"defrauded"* that will tell us if the person has been a fraud victim or not.

3.1 Target variable creation

Our dataset in itself, as mentioned in the dataset description section, does not contain a direct question that we can use to check whether a person was a fraud victim. We have to infer

this possibility from three binary questions that were present, which are *"qprod4_1"*, *"qprod4_2"* and *"qprod4_3"*. The main line of reasoning to get the value of our target variable is the following: *"qprod4_2"* and *"qprod4_3"* refer to cases where it is undeniable the respondent has been defrauded. If the reply was a solid "yes" to either, our target variable should be a one; in fact, *"qprod4_2"* refers to having private financial information stolen trough fraudulent phone calls and *"qprod4_3"* instead refers to respondent money being used to pay for goods without authorisation. Instead, *"qprod4_1"* is less clear: if say a person accepts investment advice on some worthless financial product from one of its relatives, it could be that they were in good faith and they too lost money; in a sense they were the fraud victim while the respondent was "collateral damage".

To approach this problem thus this insight might not be enough. In fact, when we consider that the possible replies for all three questions were "yes", "no" and "I prefer not to answer/I don't understand/I don't know", we see we have too many options; but considering that all three options that are not the classic "yes" or "no" give zero value to our analysis we can group them in a general "no information" of some sort. This way, we have 3^3 combinations of possible cases in which we subdivide our space of answer: "yes", "no" and "no information". Within this space, the reasoning that "if we replied yes to either second or third question then the target variable is 1", already covers 15 cases; then, we have those who were surely not victims having replied 3 times "no" as another covered possibility.

We can furthermore notice that in 4 left possibilities the respondent replied "yes" to the first question and either "no" or "no info" to the other two; in those cases we can check the *"qprod3_8"*, *"qprod3_9"* and *"qprod3_10"* questions: if the respondent replied "yes" to either of the three we can consider that whoever suggested any financial product that was later worthless was someone that had the respondent's best interest in mind and as such he was not defrauded. Else, he was; and so we are left with 7 cases where it is unclear.

Of those cases, we decided to flag as not fraud victims those where the replies where mostly "no" ("no" for two questions, "no info" for just one) and to instead delete all those rows where we essentially have not enough information because the majority of replies were neither "yes" nor "no". This drastic decision leads to loss of 110 data points that for our analysis can't be utilised, but gives us a completely partitioned space of data points where each data point is mapped accordingly to what said before to either 1 or 0.

3.2 Feature reduction

At this point we can proceed and delete some useless columns, as well as those columns we just used to create the new target variable. The other columns that brought no value and were deleted were:

- *ID*: this column is essentially useless
- *PESOFITC*: this column too is useless since we have no knowledge of how it was assigned nor what it means
- *SM*: this column give us no useful information
- *qd5b*: the number of household members should not be correlated with our target variable whatsoever, and hence removing it we actually remove the possibility of introducing biases
- *qd12*: same explanation as *qd5b*
- *qf3_99*: this is a single column of a multiple choice answer and it is eliminated because it is redundant, in fact the column "no answer" to multiple choice question is the same as all zeros for the other options from an absence of information standpoint
- *qf9_99*: same reasoning as *qf3_99*

- *qprod1c_99*: same reasoning as *qf3_99*
- *qprod3_99:* same as *qf3_99*
- *qf12_97* and *qf12_99*: same as *qf3_99*
- *qprod1_d*: once again this is somewhat redundant as our target variable shouldn't be affected by when a financial product was bought

After this feature reduction we are left with 92 columns.

3.3 Data discretisation

Another thing we can do is discretise some of the data to lower even further the number of columns and/or the width of the space of data points. As an example, it makes sense to discretise age in age ranges for both explainability and to lower the width of the space of data points. Values affected by this are:

- *AREA5*: discretisation groups values in North/center and South + islands
- *qd7*: this is exactly the age ranges creation, following standard 10 year long ranges
- *qd9*: here we are grouping education in highest *completed* education
- *qf4*: discretisation here maps answers of "I don't know" and "I don't have a personal income" together, because they both represent absence of data
- *qf8*: same as *qf4*
- *qf10_i* $\forall i$: same as *qf4*
- *qk1*: same as *qf4*
- *qf13*: same as *qf4*; here we also group the first two options together
- *qf9_i* $\forall i$: here we group together some multiple choices; we group together *qf9_2* with *qf9_3*, *qf9_7* with *qf9_8*, *qf9_1* with *qf9_9*. This allows us to reduce further the number of columns while preserving data for two columns that were very similar as couples
- *qprod3_i* $\forall i$: in similar fashion as *qf9_i*, we grouped *qprod3_1* with *qprod3_16*, *qprod3_13* with *qprod3_15*, *qprod3_3* with *qprod3_6* and *qprod3_5* with *qprod3_7*, *qprod3_12* and *qprod3_14*
- *qf12_i* $\forall i$: here in similar fashion as *qf9_i* we aggregate *qf12_1_c* with *qf12_3_g* and *qf12_5_m* with *qf12_4_k*, *qf12_5_o* and *qf12_6_p*
- *qk3* to *qk7* series questions: these questions are related to financial knowledge and as such they can be viewed as a three-class correct/incorrect/blank answer; hence the discretisation criteria is exactly that

3.4 Missing values handling

One more manipulation that we must do is handle missing data in our dataset. This refers specifically to columns generated by questions *"qf12_i"* series, *"qprod2"* and *"qprod3_i"* series, which are only asked to the respondent when he replies in a specific manner to earlier questions. This time the handling is really simple: basically whenever we see a NaN we substitute it with whatever value in that question is associated with the absence of a reply or information (usually -99). Notice how once again this has basically no impact on the data itself and is consistent and coherent. Now, after all this work, as is the dataset is comprised of 2266 rows of data by 79 columns.

4 Feature engineering

Now, while our dataset is technically complete and without missing data, it still is far from easy to interpret, even despite the light aggregation on practically identical columns. We will utilise the dataset as is to train a model, but we would also like to be able to interpret the model more easily, and that can be done through feature engineering.

More specifically the idea is the following: we want to exploit common financial clichés to check whether they can actually hold information and whether the "fraudability" of each participant can be measured through naïve features with which most people would judge one's financial situation. This way, if machine learning models trained on those feature have similar results as those trained on the full dataset we will not have loss of information or predictive power, while having a great explainability and of easy interpretation, linked to common thought; if this is not the case, we will be anyway able to tell if that's because those naïve features are not sophisticated enough or if it's just that frauds occur at any level without any correlation to new engineered variables; effectively, we'll be able to gain knowledge and interpretability either way and in any case the result would provide guidance to any regulatory authority interested in protecting citizens.

To transform existing features in new ones we decided to use a quasi linear combination of variables, which is described by the matrix 10.2; the combination is almost a linear combination but sometimes we decided to either transform variables such that they were centered around zero or enhance their negative effect if they were binary. This means for example that all questions with ranges 1 to 5 (*"qf10_i"* series, *"qk1"*) were rescaled with a $\bar{x} - x_i$ transformation, and then that value was used in the linear transformation, or another example of this is the fact that for some features some instances of *"qf12_i"* or *"qf9_i"* were grouped such that even if only one was flagged as a "yes" then the reply on the others didn't matter. The transformation because of this details isn't entirely linear aside for some features (as an example for "*financial knowledge*").

The new features are:

- ***Financial knowledge***

  The feature incorporates the survey questions that estimate respondents' financial knowledge. Specifically, these questions test financial knowledge through targeted mathematical/financial questions: respondents can enter an exact value from the keyboard or choose one of the possible answers from those reported. The minimum theoretical possible value corresponds to -7 and the maximum theoretical possible value corresponds to 7; this range of possible results for this feature was obtained by calculating the worst and best case of the transformation. Note that this score has nothing to do with perceived financial knowledge, which will be defined afterwards, at least in principle

- ***Indebtment***

  The debtor variable shows whether the respondent of the survey leans towards debt or is debt averse. High values correspond to a tendency towards debt, while low values correspond to a debt averse person. The features that are included in this variable are related to different aspects of the financial situation of each respondent such as missing payments, worrying about normal living expenses or signing up for unsecured loan. To better understand the variable we are dealing with, we compute the lowest value in the dataset, which corresponds to -5.5, and highest value which is equal to 9.0. More precisely, we can draw our own conclusions using the theoretical minimum and maximum. The range for this feature is from -5.5 to 10.5, calculated considering the highest possible case and the lowest one of the transformation. The discrepancy between theoretical maximum and our dataset's maximum score already suggests a general debt-aversion.

- ***Perceived financial knowledge***

  The perceived financial knowledge tells us how much the respondent felt confident in its financial knowledge: the higher, the higher the perceived knowledge. The lowest value and the highest one within the dataset are equal to -5.25 and 6.0, which are very closed to the theoretical values of the variable that are -6.0 and 6.5. The variable is computed with questions such as *"qk1"*.

- ***Saving attitude***

  The saving attitude variable tells how much the respondent is a saver or someone who tends to spend: the higher, the more the respondent saves. More in depth, this feature takes into account all those aspects of real life where a person is willing to spend or save money, through questions on both the actual savings and the person put in hypothetical scenarios. The lowest and highest value calculated within the dataset are equal to -6.25 and 10.85. In this case, the range of values is quite wide, which is confirmed by the theoretical minimum and maximum values of -7.75 and 15.85.

- ***Planning attitude***

  The planning attitude feature highlights how much the respondent plans his future or not: the higher, the more the respondent usually is able to plan for a distant (or not) future, financially speaking. Tendency to live for today and letting tomorrow take care of itself or setting long term financial goals are only two examples of what is included as planner attitude. In this cases, the lowest value within the dataset is equal to the theoretical minimum which corresponds to -8.0, while the highest value is 10.75 and the theoretical maximum is 11.75.

- ***Financial products experience***

  The financial products experience feature wants to measure respondents' knowledge and experience of various financial products (such as bonds, stocks, shares, mutual funds). Specifically, the questions tried to test direct financial experience. The lowest value obtained among the respondents corresponds to -0.5, while the highest was observed to be 8.5; finally, the range of possible results for this feature was obtained by calculating the maximum and minimum of the transformation and yielded a value of -0.5 for the minimum and and the 10.6 for the maximum.

- ***Digitalisation***

  The digitalisation feature incorporates a series of questions designed to better understand the relationship between respondents and technology. In particular, these questions assess the use of technological or analog products (such as cash) within the domain of financial economics. It was observed that the lowest value obtained among the respondents was -2.2, while the highest was 1.0, while the range of possible results for this feature was obtained by calculating the maximum and minimum of the transformation: the minimum theoretical value is -3.0 and the maximum theoretical possible value is 1.0.

- ***Independent financial approach***

  The independent financial approach variable tells how much the respondent has a tendency in being independent around financial matters: the higher, the more the approach is independent. This characteristic considers not only independence from an economic point of view, where the emphasis is on the ability to cover living costs, but also independence from the point of view of making financial decisions, where advice from friends/relatives or independent financial advisers may or may not be relevant. We compute the lowest value for this feature which is -4.5 and the highest one which is 5.5. The theoretical minimum and maximum that define the range for independence financial approach are -9.0 and 7.0.

There are some features that unfortunately we were not able to include in this creation, namely: *"qf9_10"*, *"qprod3_18"*, *"qf12_7_r"*. These features were relevant because in some multiple choice question they represented the "other" option, but that clearly does not bring any information for us to be able to aggregate it in any of the new features. As such, even though we know we could introduce some slightly incorrect information in the dataset we still decided to drop them as the number of replies where it could introduce noise was low enough.

5 Descriptive statistics

We initially examined the correlation matrix and the partial correlation matrix on the dataset with the new features engineered. The two correlation matrices provide an intriguing point of departure for analyzing the relationships between the variables in the dataset from two complementary perspectives. The relationships between the variables are represented with their correlation coefficients; in the case of the simple correlation matrix, the degree of linear association between two variables is observed. In contrast, the partial correlation matrix allows for the measurement of the association between two variables after the effect of other variables in the dataset has been removed. This approach helps to exclude any potential indirect or spurious correlations.

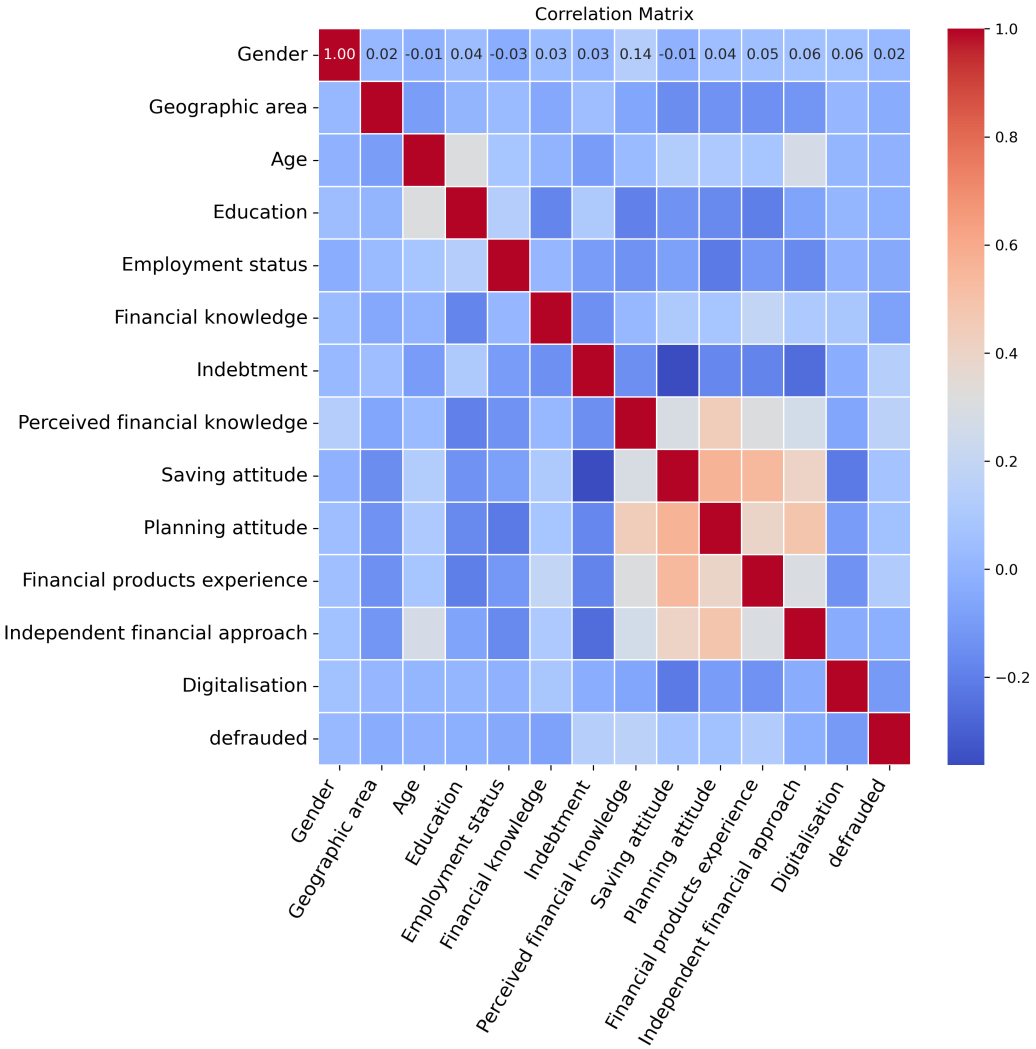

Fig. 1: Correlation matrix

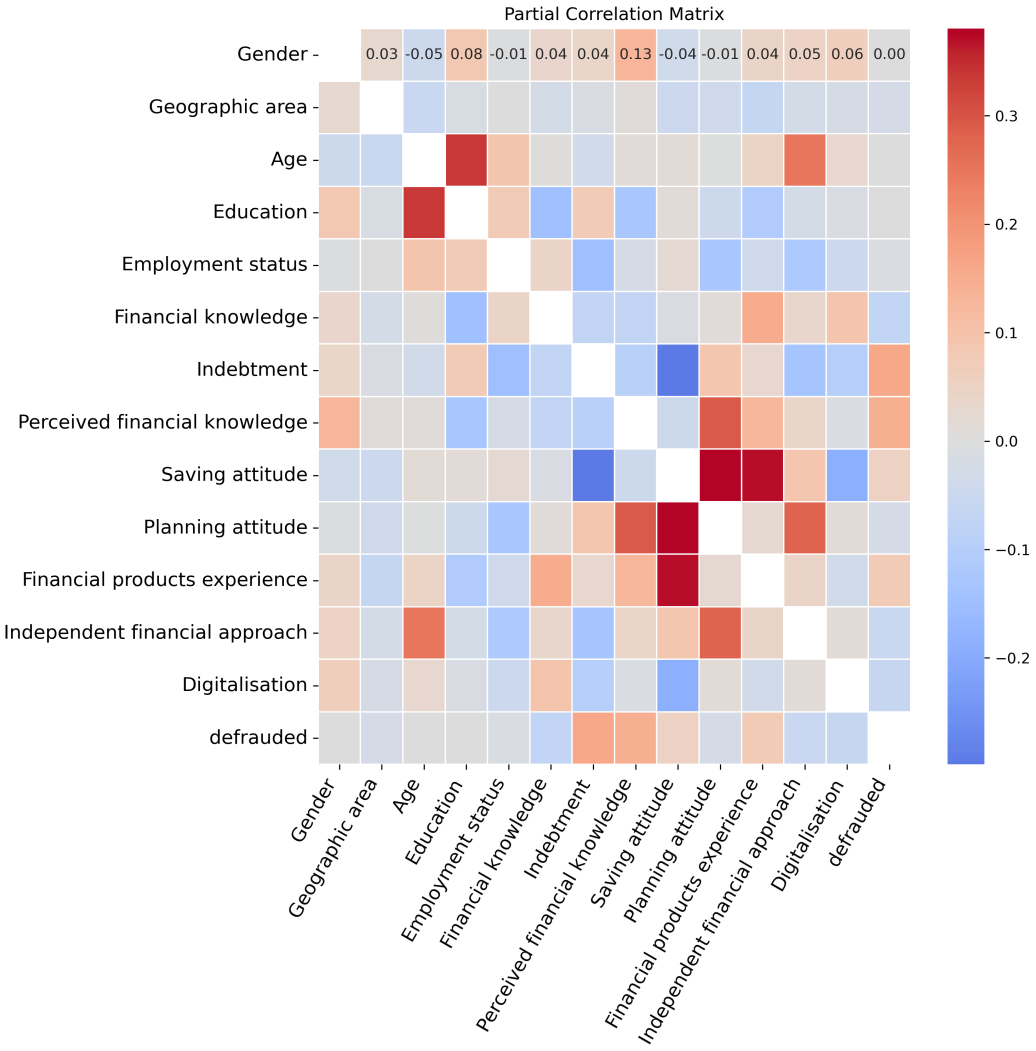

Fig. 2: Partial correlation matrix

Upon examination of the correlation matrices, it becomes evident that the main diagonal of the correlation matrix [Fig.1] is consistently equal to 1, as each variable is perfectly correlated with itself; in contrast, the partial correlation matrix [Fig.2] exhibits a diagonal with values equal to 0, which is a consequence of the removal of all its influence with itself. Furthermore, an area of distinction can be observed in the correlation matrix [Fig.1], differing in color from the overall matrix and representing the correlation between some of the features generated through feature engineering. It can be observed that the relationship between the features, which were created through the aforementioned process, is relatively neutral and generally more decorrelated than the others. This result demonstrates that the feature engineering process facilitated the generation of new features that are well-structured, independent of one another and highly informative with regard to the analysis context. In contrast, the partial correlation matrix [Fig.2] indicates a strong relationship between:

- "*Planning attitude*" and "*Saving attitude*" which can be explained with the fact that saving usually implies planning and vice versa to a certain degree
- "*Financial products experience*" and "*Saving attitude*" which can be explained thanks to the fact that to save in the digital age one is somewhat forced to open and use financial products
- "*Education*" and "*Age*" which we can interpret through the lens that generally speaking

younger people may have not yet completed their education

- "*Saving attitude*" and "*Indebtment*" which we can explain thinking that in general being more of a spender often implies having more debt and having more savings implies generally speaking not having to resort to debt when facing difficulties

6 Data Analysis

The most versatile and most useful classifiers of choice for our project are tree based models and the logistic regression, which keep explainability and are generally versatile in predicting a binary output, which in our case corresponds to predict if the respondent of the questionnaire has been a victim of a financial fraud. To be more precise, we decided to test three different classifiers:

- *Random Forest*, which is a collective learning method that exploits a committee of decision trees
- *Logistic Regression*, which uses parametric conditional probability to assign new observations to one of the target's class
- *Multi-Layer Perceptron*, which is a specific case of the *Artificial Neural Network*, composed by multiple hidden layers

We train these three classifiers based on the *Matthews Correlation Coefficient* (MCC) [4] [5] as a measure of the quality of binary classification. The MCC takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used if the classes are of very different size, which is the case in our dataset. The MCC returns a value between -1 and +1 where +1 represents a perfect correlation between predicted and actual values, 0 no better than random prediction and -1 complete negative correlation between prediction and observation. The MCC can be directly calculated from the confusion matrix and we can find its value as:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{1}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

To get results that make more sense and are averaged over different partitions of the dataset without being case-specific, we decided to combine a cross-validation method with each classifier. The cross validation uses a different strategy to train and evaluate the analysis made; the number of k folds you can use can vary, although the value we considered optimal for our case was 10. The dataset was divided into ten subsets and the three models were trained ten times. During each iteration, nine subsets served as the training set, while one subset functioned as the validation set. The validation set in each iteration is always a different one to prevent the same scenario to be recreated; this process facilitates a more robust comparison among all precedent models.

We start our analysis doing a simple test and then we move on to implement some techniques trying solving the main problems of the dataset. In particular, the simplest thing we can do to combat imbalances across the dataset is try and balance class weights for Random Forest and Linear Regression. Moving forward, we implement more drastic changes such as undersampling with and without class weights. The last thing we do for a better results interpretation is to test the best model at which we arrive with both the whole dataset, but also the new dataset composed by the socio-demographic and the engineered features. Below we move to a better explanation with more details.

6.1 Simple test

The most simple test we can do is try and approach the problem with simply feeding a Random Forest model, a Logistic Regression model and a Multi Layer Perceptron model our cleaned dataset, without any feature engineering, and without any methods to combat the class imbalance we already know present, apart from stratification in dividing test and training set. This, as can be somewhat expected, yields very poor results: even through a 10-fold cross validation Random Forest achieves an MCC of basically zero, Multi Layer Perceptron around 0.1 and Logistic Regression around 0.17; the problem is the class imbalance, as highlighted by extremely low values of recall and precision for class 1 of our target variable in all three cases. It is interesting to notice that the overall predictive power is not bad as these metrics, given that the AUC for Random Forest is highest (around 0.84), while for Logistic Regression and Multi Layer Perceptron decreases to around 0.71 and 0.64.

6.2 Class weights

One of the first approaches we can do, at least in the case of Random Forest and of Logistic Regression, is to train the model such that class 1 is way more important that class 0: the model is penalised more when making mistakes on class 1, and hence should learn more. We decided to keep a proportion of classes similar to the proportion of data in our dataset, that is penalising errors on class 1 twelve times more than what is done on class 0.

This approach is not possible on Multi Layer Perceptron because of its structure, but it's not big problem, considering results do not particularly improve: we see in fact a slight rise of Random Forest to an MCC of around 0.1 and essentially no effect on Logistic Regression; moreover overall predictive power, measured trough AUC, generally stays the same as in simple test.

6.3 Undersampling with no Class Weights

Another, more drastic approach is to undersample our dataset: using a random undersampling function that essentially cuts the excess of samples of class 0 is indeed drastic, but completely eliminates by definition any imbalances. Moreover this approach is applicable to all three models we considered, as it is implemented at the train-test split level.

This time results do improve, and a lot: after the usual 10-fold cross validation, Random Forest improves to an astonishing 0.4 MCC, while Logistic Regression to a more modest 0.3 and Multi Layer Perceptron to a 0.2 value of MCC. Given the fact that the undersampling is done at the train-test split we also averaged over 50 different undersampled versions along with the usual 10-fold cross validation and we found that our scores did not change particularly, with an MCC of 0.40, 0.28 and 0.26 respectively for Random Forest, Logistic Regression and Multi Layer Perceptron. Loss of predictive power is to be expected when removing samples, and indeed AUC is less than before, but not considerably less: in fact, this can be seen in figure [Fig.3].

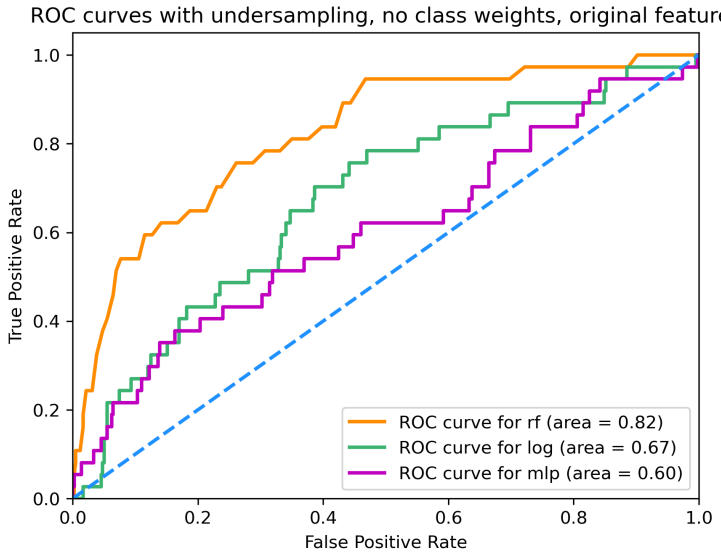

Fig. 3: ROC curves of original features with undersampling and no class weights

6.4 Undersampling with Class Weights

We can now try and combine both class weighting and undersampling; skipping already at the averaged and cross validated MCC scores we see that this time they actually do not improve, but rather they worsen from the case of just undersampling, with an MCC of 0.35, 0.13 and 0.25 respectively, signaling that uniting class weighting actually creates imbalances after the dataset had been undersampled exactly to combat imbalances. AUC is also impacted more negatively than before, as seen in figure [Fig.4]; note that clearly values for Multi Layer Perceptron stayed the same as it can't undergo class weighting.

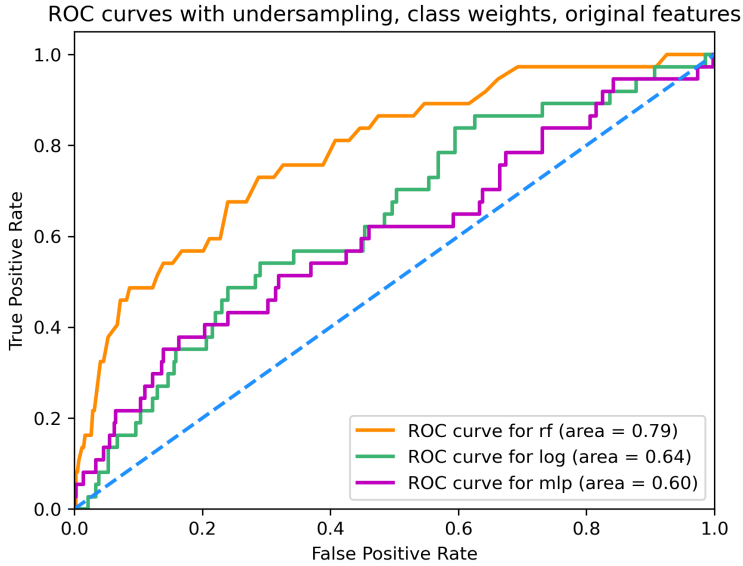

Fig. 4: ROC curves of original features with undersampling and class weights

6.5 Feature Engineering Impact

At last, we have to ask ourselves if the process of feature engineering actually worsens our MCC and/or AUC; in general we can expect a loss in AUC but MCC should not vary too much as the dimensionality of the dataset is reduced and hence data points of interest are less sparse,

albeit we lose information. We also know the new engineered features are uncorrelated with one another and so we can expect the model to not lose too much predictive power. For this analysis we'll consider only Random Forest with undersampling and no class weights, as it proved to be the best combination. With this configuration, the MCC is computed to be 0.35 when averaged over many undersampling iteration, with a slight decrease as expected, while AUC, depicted in figure [Fig.5] decreased a little to a mean value over iterations of 0.73, signaling that what was expected is found.

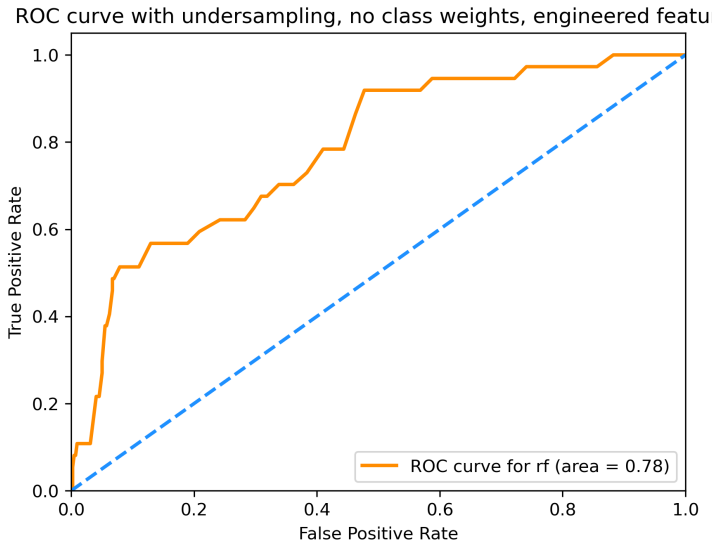

Fig. 5: ROC curves of one iteration for engineered features with undersampling, no class weights

This means that for any result we can safely use as best model Random Forest with no class weights but with undersampling and applied to the engineered dataset, because the decrease in MCC and predictive power is small compared to the ease of interpretability within the dataset. This is optimal as it provides the best combination to perform feature selection while having an high level of interpretability.

7 Results

The first observation we do is related to the value of the MCC. In fact, one could argue that even a value of 0.35, which was the best we obtained, is not particularly high and that's true; but this has to be balanced with some arguments that must be taken into consideration.

First things first, we know people are not always entirely truthful and/or precise; this is also especially true in Italy, where financial matters remain somewhat shrouded in a cloak of privacy and most of the time people feel attacked when asked questions in this sphere. This means that there will always be some systematic error given by these inconsistencies, and this error resulted at the start of the dataset in us losing some data points, but also later it can introduce quite some noise. This means an MCC close to 1 may never be obtained, and we have to aim lower; estimating the effects of this noise is an even harder task though.

Then, the imbalance of the dataset and its high dimensionality definetly do not help. Computing some more metrics on our latest Random Forest model we see that we have, for our class of interest, values of 0.66, 0.68 and 0.65 for mean F1 score, mean precision and mean recall. These values indicate that generally speaking we have achieved a good tradeoff with predictive power and interpretability and explainability, which was what we were aiming for in the first place.

At last, to some degree we also have to probably recognise that being fraud victims, contrary

to popular belief, is an element of society that is transversal, that is can happen to everyone, and everyone has to be ready and prepared for when the time comes.

We now want to aim at identifying any pattern that might be linked to being victims of a financial fraud in the contest of the stereotypical features we engineered and to do so we start with a simple feature importance with Gini entropy decrease. Since it does not take into account collinearity within features, we double check using Permutation Importance and compare the results. Analysed the importance of each feature, we investigate the impact each feature has on the target variable with Partial Dependences evaluating whether the findings align with the collective imagination or deviate from conventional wisdom. Lastly, we try to create some stereotypical profiles of people who might have replied to our survey to check what is the probability of them to being defrauded to compare it to common thought and to provide a somewhat practical example.

All these analyses will be done on a Random Forest model with undersampling, no class weights, trained on the engineered features' dataset. More details on the results found are described in the following sections.

7.1 Feature Importance

The first analysis that is very classical and can be done is looking at the importance of each feature through the Gini index; this method is based on the decrease in impurity (e.g., Gini impurity or entropy) that each feature causes when used to split a node in the decision trees. The importance of a feature is calculated as the total reduction in the impurity across all trees in the forest, averaged over all trees. This is what is present in figure [Fig.6].

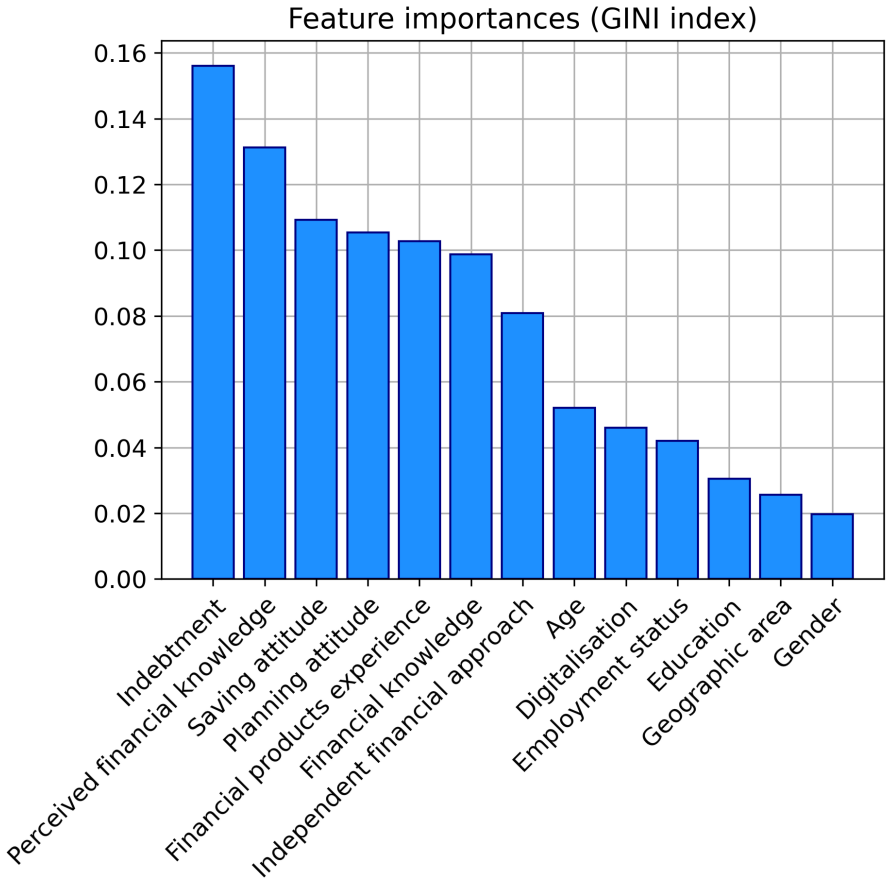

Fig. 6: Feature importance considering Gini entropy decrease

It's interesting to notice how, for our model, the most important attributes were all of the financial type, and not socio-demographic characteristics or tendencies such as digitalisation approach.

Now to double check we also take a look at feature importance through permutation, in figure [Fig.7].

Permutation importance works in a fundamentally different way than Gini importance. The idea is that per each feature it shuffles its values in the dataset. This breaks the association between that feature and the target variable while keeping the other features the same; the idea is to simulate a situation where the feature provides no useful information to the model. Then it re-evaluates the model on the same test set; the drop in performance is calculated, indicating how important the feature was. A large drop in performance means the feature was important, while a small or no drop means the feature was not important.

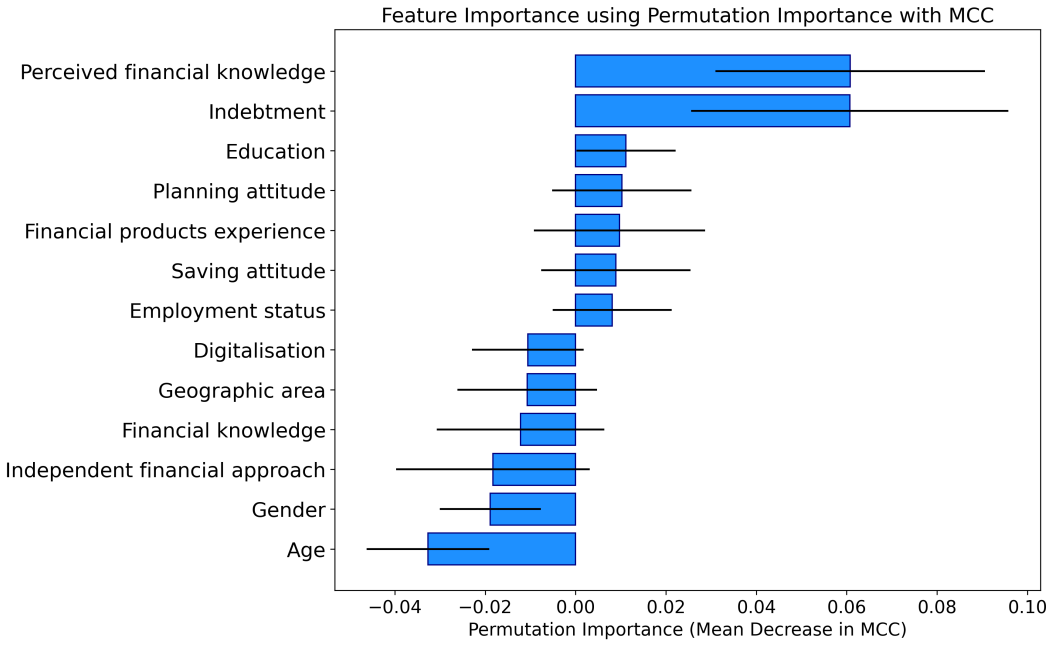

Fig. 7: Feature importance considering feature permutation

This time we can see that the two most important features were, within error bars, *"Perceived financial knowledge"* and *"Indebtment"*, which were also the top two in Gini importance; conversely, we once again see that socio-demographic variables such as *"Age"* and *"Gender"* are not important; we also see that probably *"Financial knowledge"* and *"Independent financial approach"* exhibit collinearity with other variables, given that their specific importance is lower than what was present in the Gini plot.

7.2 Feature Partial Dependence with Target Variable

As a more in depth analysis, we evaluate the influence of each characteristic on the target variable, taking partial dependencies into account. This will enable us to ascertain whether the results are in line with the collective imagination or deviate from conventional wisdom. The plots in figure [Fig.8] show the 13 possible combination between each feature and the defrauded feature.

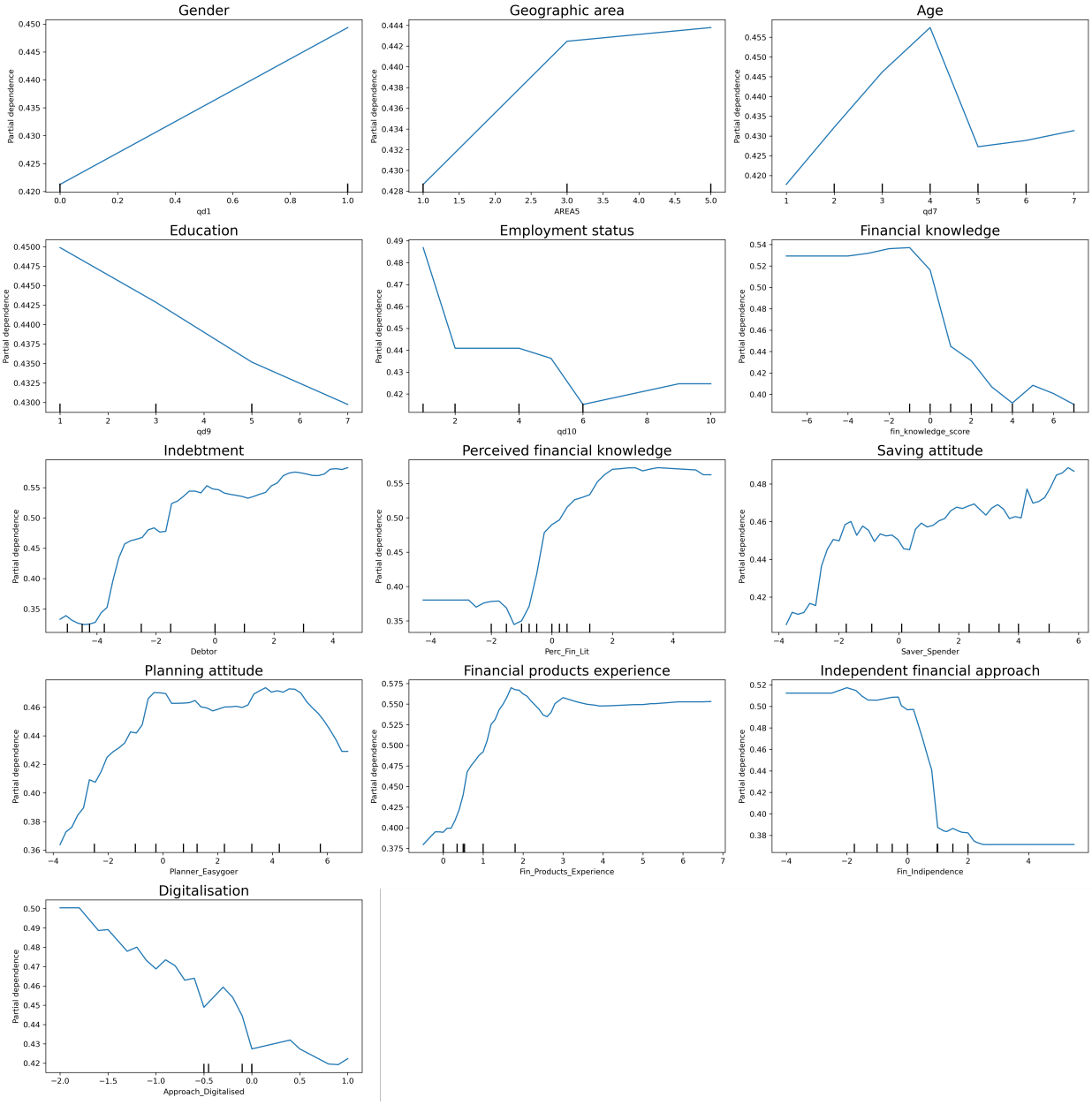

Fig. 8: Feature partial dependencies with target variable

What can be seen here that is peculiar is how quickly the chances of being fraud victims rise with *"Perceived financial knowledge"* and *"Indebtment"*, and also with *"Financial products experience"*. A surprise to be sure is instead the relationship that seem to arise between being a fraud victim and *"Saving attitude"*: it looks like often times those who are defrauded also display a tendency on saving money and not spending it. This is probably due to the fact that it might be the case that those who are not interested in actively saving actually can't get hooked in financial frauds under the promise of a quick buck or more returns on investments.

From our research question point of view, it's plots like the one for *"Financial knowledge"* that show us how to combat fragility from a financial security standpoint in our society: by enhancing the financial knowledge of citizens. As soon as it reaches a basic level there is in fact a big drop in the dependence.

7.3 Units Out of Sample

We can now try and take a look at different profiles which are clearly out of sample, but instructive because they might be examples of stereotypical people we meet or we can see ourselves into.

Features	**Profile 1**	**Profile 2**	**Profile 3**	**Profile 4**	**Profile 5**	**Profile 6**
Gender	1	0	1	0	1	0
Geographic area	1	1	5	3	1	3
Age	4	1	3	7	5	2
Education	1	1	5	7	3	1
Employment status	2	9	5	6	1	2
Financial knowledge	-4	-5	-6	-4	-7	7
Indebtment	-4	-5	3	-5	8	-1
Perceived financial knowledge	1	-2	5	-4	6	6
Saving attitude	11	-4	-5	12	2	7
Planning attitude	7	4	-4	-6	0	8
Financial products experience	5	1	1	6	3	8
Independent financial approach	5	-5	3	6	4	7
Digitalisation	-2	1	-1	-3	0	1

Table 1: Six different profile creations

- **Profile 1**: male adult aged 45-54 from Northern Italy, in full-time employment and with a university degree. Despite lacking financial expertise, he considers himself to possess a considerable amount of knowledge in this area and relies on his own understanding. He has a medium to high level of wealth, enabling him to have savings and no debts. He is reluctant to embrace technology

- **Profile 2**: female aged 18-25, from Northern Italy. She is a student with a bachelor's degree and is aware that her financial knowledge and experience with financial products are limited. Consequently, she seeks guidance from experts in these fields. Due to her lack of income, her financial resources are constrained, making it challenging for her to save and set ambitious long-term financial goals. However, she demonstrates a high level of digital proficiency

- **Profile 3**: male adult aged 35-44 from Southern Italy, seeking employment and possessing a primary school diploma. He considers himself to be knowledgeable about finance and therefore relies on his own expertise, however, evidence suggests otherwise. Lacking a stable income, he has limited financial resources, resulting in a lack of savings and an accumulation of debt that impedes his ability to plan for a secure future. Additionally, he exhibits a lack of confidence in technology

- **Profile 4**: elderly female, aged 75 and over, from central Italy, who has been retired for a number of years. She has some primary school certificate and, given her age, is aware that she lacks confidence in financial matters. However, she has gained some experience of financial products over the course of her life and, as a result, rarely seeks advice from family members or experts. As she is retired, she has no debts and saves money on small things, as she is unable to plan for the distant future. She has a negative attitude towards technology

- **Profile 5**: male adult aged 55-64 years old from Northern Italy. He is an entrepreneur with a high school diploma. He considers himself to be knowledgeable about finance and therefore relies on his own expertise, which is frequently demonstrated to be erroneous.

His employment provides a satisfactory income, but he is unable to save due to the burden of multiple debts. He attempts to remain current with new technologies

- **Profile 6**: woman aged 25-34, hailing from central Italy. She is gainfully employed in a banking position and holds a university degree. She possesses exemplary financial acumen and experience, which she deems indispensable. She allocates a portion of her income to savings, and her financial obligations include a mortgage and regular installments, which are typical of individuals of her age. She is highly proficient in digital technologies

The predicted probabilities for these out of sample units can be seen in figure [Fig.9]:

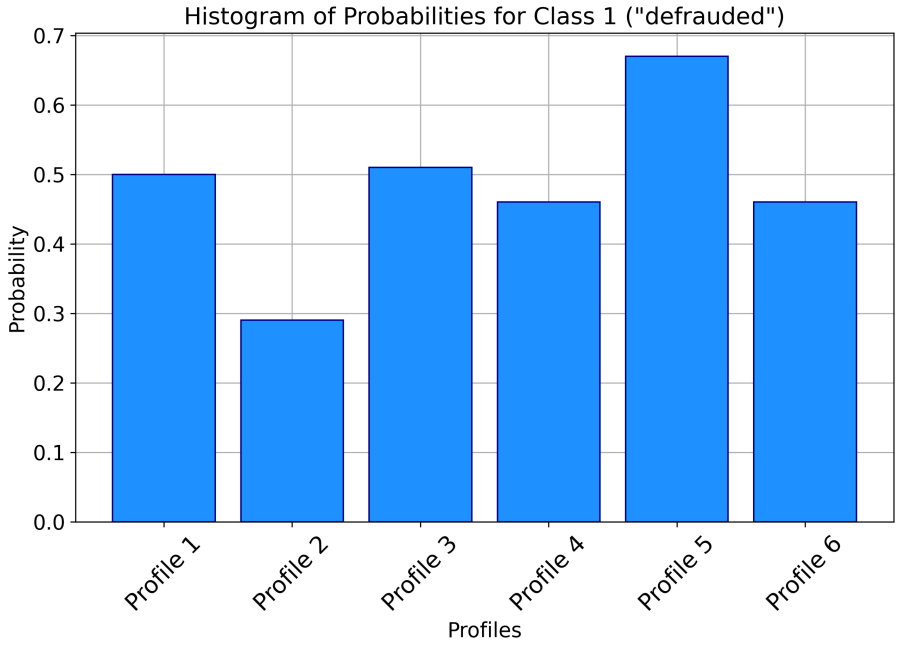

Fig. 9: Probabilities of six stereotypical profiles being defrauded

While most profiles are close to 0.5 and hence somewhat uncertain, we have two profiles that stand out quite a lot, one negatively (Profile 2) and one positively (Profile 5); we can see that the one that stood out negatively already had debt and perceived himself as knowledgeable about financial matters, while the one that stood out positively was generally debt averse and knew about its limitations around financial knowledge. Once again these two prove to be the most important factors.

Another thing this plot shows is that profiles that we might think are not vulnerable, such as Profile 6, at last indeed seem quite vulnerable, and to some level being fraud victims is a transversal element of our society, sadly.

8 Conclusion

To conclude, does common knowledge identify fragile subjects? From our research the answer seems to be only partially yes, and for some other aspects no. For example, it is highlighted how fraud is a transversal aspect within the digital age; and moreover, not all common knowledge seems to be pointing out in the correct direction, such as in the case of partial dependencies for the saving attitude variable or in the case of Profile 6 of our out of sample profiles, in a manner that could be linked to the conjunction fallacy (also known as the "Linda problem") [6]. But still some aspects of common knowledge seem to emerge as correct from the inherent noise of the dataset: for example, the fact that being in debt makes someone much more vulnerable than

others towards frauds, or also the matter of financial knowledge, which seems to be omnipresent and linked to many variables within the analysis. Especially regarding financial knowledge, it is interesting how it is usually pointed out as one of the first defense barriers against financial frauds, but it is peculiar how it is so not because of the information that it conveys to the one who possesses it but rather because it can disprove one's fallacies around the financial world, both in them being too sure of themselves and their behaviour and attitudes towards savings, debt, financial products and expectations of returns.

At last, this research highlights once again another important albeit reiterated aspect: that fragile subjects are uniformly distributed across socio-demographic classes, meaning there is no particular correlation that is significant for sex, age, education or even digitalisation.

This all means that any act that wants to reduce the vulnerability of the Italian population towards financial frauds has to act across many socio-demographic classes, with a particular accent on attitudes but also financial knowledge and also has to try and build measures and protective barriers around already financially stressed individuals. As always in a sense, complex problems cannot be reduced to simple and quick solutions but have to be thoroughly analyzed and require custom made solutions.

9 Improvements

Further improvements to this type of research are plenty, as an example we have:

- the more data available the better: given other datasets with about the same imbalance (or a better one), being able to aggregate their data in the engineered features would mean being able to access new data points which can highlight new relationships; this presents many challenges that include, but are not limited to, the choice of how to reduce features to engineered ones; the ability to recreate effectively the target variable; the problem of noise in each dataset

- other approaches: approaches that might trade some interpretability for more complex models might work better in predicting if a person is more fragile and could ultimately lead to other insights; an example of such an approach could be clustering, or more complex ML models

- a different choice of data aggregation: from varying weights to considering different and less linear transformations to even changing the interaction between terms of the combinations, a different choice of data aggregation could lead to different insights; there would still be some systematic problems such as noise in the dataset because of inconsistencies but some problems could be resolved this way

- a more practical approach: rebuilding of the original dataset with new participants, maybe found through the internet or means where more anonymity is present, could bring about more data, or data that is best suited for the research question, or could even lead to create some tools that anyone can use to assess its fragility in the space of financial frauds

10 Appendices

10.1 Questionnaire used

The questionnaire consists of the following questions which correspond to columns of the analysed dataset:

- ***ID***: Individual identifier
- ***SM***: Model of interview (0 - Tablet; 1 - CAPI)
- ***PESOFITC***: Sample weights
- ***qd1***: Gender (0 - Female; 1 - Male)
- ***AREA5***: Geographical area (1 - North-West; 2 - North-East; 3 - Centre; 4 - South; 5 - Islands)
- ***qd5b***: Number of household members (1 - One; 2 - Two; 3 - Three; 4 - Four; 5 - Five; 6 - Six or more)
- ***qd7***: Age
- ***qd9***: Educational qualification (1 - University-level education; 3 - Complete secondary school; 4 - Some secondary school; 5 - Complete primary school; 6 - Some primary school; 7 - No formal education)
- ***qd10***: Employment status (1 - Self-employed; 2 - In paid employment; 4 - Looking after the home; 5 - Looking for work; 6 - Retired; 9 - Student; 10 - Other)
- ***qd12***: Country of birth (1 - Italy; 0 - Abroad)
- ***qf1***: Who is responsible for making day-to-day decisions about money in your household? (1 - You make these decisions by yourself; 2 - You make these decisions with other household members; 3 - Other household members, not you; -99 - No answer)
- ***qf2***: The next question is about the household budget. A household budget is used to plan what share of your household income will be used for spending or saving. Does your household have a budget to plan in advance which share of income will be used for spending and which share will be saved for the next years? (1 - Yes; 0 - No; -99 - No answer)
- ***qf3_1 - qf3_99***: In the past 12 months have you been personally saving money in any of the following ways, whether or not you still have the money? Please do not include pension savings.
  - ***qf3_1***: Saving cash at home or in your wallet
  - ***qf3_3***: Paying money into a savings account
  - ***qf3_4***: Giving money to family to save on your behalf
  - ***qf3_6***: Buying financial investment products, other than pension funds (e.g. bonds, shares, mutual funds)
  - ***qf3_7***: Or in some other way (including remittances, buying livestock, gold or property)
  - ***qf3_8***: Has not been actively saving
  - ***qf3_99***: No answer

- ***qf4***: And if you, personally, faced a major expense today, equivalent to your own monthly income, would you be able to pay it without borrowing the money or asking family or friends to help? (1 - Yes; 0 - No; -97 - Don't know; -98 - I don't have a personal income; -99 - No answer)

- ***qf8***: Overall, how confident are you that you have done a good job of making financial plans for your retirement? (1 - Very confident; 2 - Confident; 3 - Somehow confident; 4 - Not very confident; 5 - Not at all confident; 6 - I'm not planning for retirement; -97 - Don't know; -99 - No answer)

- ***qf9_1 - qf9_99***: And how will you - or do you - fund your retirement?
  - ***qf9_1***: From drawing a government pension/old-age benefit
  - ***qf9_2***: From an occupational or workplace pension plan
  - ***qf9_3***: From a private pension plan
  - ***qf9_4***: From selling your financial assets (such as stocks, bonds or mutual funds)
  - ***qf9_5***: From selling your non-financial assets (such as a car, property, art, jewels, antiques)
  - ***qf9_6***: From income generated by your financial or non-financial assets
  - ***qf9_7***: By relying on a spouse or partner to support you
  - ***qf9_8***: By relying on your children or other family members to support you
  - ***qf9_9***: Survivor's pension
  - ***qf9_10***: Other
  - ***qf9_99***: No answer

- ***qprod1c_1 - qprod1c_99***: In the last two years, which of the following types of financial products have you bought, whether or not you still hold them?
  - ***qprod1c_1***: A pension or retirement product
  - ***qprod1c_2***: An investment account such as a unit trust
  - ***qprod1c_3***: A mortgage or a bank loan secured on a property
  - ***qprod1c_5***: An unsecured bank loan or a salary/pension-backed loan
  - ***qprod1c_6***: A credit card
  - ***qprod1c_7***: A current/checking account
  - ***qprod1c_8***: A savings account
  - ***qprod1c_10***: Insurance
  - ***qprod1c_11***: Stocks and shares
  - ***qprod1c_12***: Bonds
  - ***qprod1c_14***: A prepaid debit card/payment card
  - ***qprod1c_99***: No answer

- ***qprod1_d***: Which of these did you choose most recently? (If at least one product was selected go to *prod1_d*, otherwise go to *qf10*)

- ***qprod2***: Which of the following statements best describes how you made your choice? (1 - I considered several options from different companies before making my decision; 2 - I considered the various options from one company; 3 - I didn't consider any other options at all; 4 - I looked around but there were no other options to consider; -99 - No answer)

- ***qprod3_1 - qprod3_99***: Which sources of information do you feel most influenced your decision?
  - ***qprod3_1***: Unsolicited information sent through the post
  - ***qprod3_2***: Information picked up in a branch
  - ***qprod3_3***: Product specific information found on the internet
  - ***qprod3_4***: Information from sales staff of the firm providing the products
  - ***qprod3_5***: Best-buy tables in financial pages of newspapers/magazines
  - ***qprod3_6***: Best-buy information found on the internet
  - ***qprod3_7***: Specialist magazines/publications
  - ***qprod3_8***: Recommendation from independent financial adviser or broker
  - ***qprod3_9***: Advice of friends/relatives (not working in the financial services industry)
  - ***qprod3_10***: Advice of friends/relatives (who work in the financial services industry)
  - ***qprod3_11***: Employer's advice
  - ***qprod3_12***: Newspaper articles
  - ***qprod3_13***: Television or radio programs
  - ***qprod3_14***: Newspaper adverts
  - ***qprod3_15***: Television adverts
  - ***qprod3_16***: Other advertising
  - ***qprod3_17***: My own previous experience
  - ***qprod3_18***: Other source
  - ***qprod3_99***: No answer
- ***qf10_1 - qf10_12***: I am now going to read out some statements. I would like to know how much you agree or disagree that each of the statements applies to you, personally. Please use a scale of 1 to 5, where 1 tells me that you completely agree that the statement describes you, while 5 shows that you completely disagree (1 = Totally agrees, 5 = Totally disagrees, -97 = Don't know, -99 = No answer)
  - ***qf10_1***: Before I buy something I carefully consider whether I can afford it
  - ***qf10_2***: I tend to live for today and let tomorrow take care of itself
  - ***qf10_3***: I find it more satisfying to spend money than to save it for the long term
  - ***qf10_4***: I pay my bills on time
  - ***qf10_5***: I am prepared to risk some of my own money when saving or making an investment
  - ***qf10_6***: I keep a close personal watch on my financial affairs
  - ***qf10_7***: I set long-term financial goals and strive to achieve them
  - ***qf10_8***: Money is there to be spent
  - ***qf10_9***: My financial situation limits my ability to do the things that are important to me
  - ***qf10_10***: I tend to worry about paying my normal living expenses
  - ***qf10_11***: I have too much debt right now
  - ***qf10_12***: I am satisfied with my present financial situation

- ***qprod4_1 - qprod4_3***: Thinking about financial products and services in general, in the last 2 years, have you experienced any of the following issues? (1 = True, 0 = False, -95 = I don't understand, -97 = I don't know, -99 = I prefer not to answer)
  - ***qprod4_1***: You accepted advice to invest in a financial product that you later found to be worthless
  - ***qprod4_2***: You accidentally provided financial information in response to an email or phone call that you later found out was not genuine
  - ***qprod4_3***: You discovered that someone had used your cards or personal information to pay for goods without your authorization
- ***qk1***: And now we talk about financial literacy. Could you tell me how you would rate your overall knowledge about financial matters? (1 - Well above average; 2 - Above average; 3 - Average; 4 - Below average; 5 - Well below average; -97 - Don't know; -99 - No answer)
- ***qf11***: Sometimes people find that their income does not quite cover their living costs. In the last 12 months, has this happened to you, personally? (1 - Yes; 0 - No; -99 - No answer)
- ***qf12_1_a - qf12_99***: What did you do to make ends meet the last time this happened?
  - ***qf12_1_a***: Draw money out of savings
  - ***qf12_1_b***: Cut back on spending, spend less, do without
  - ***qf12_1_c***: Sell something that you own
  - ***qf12_2_d***: Work overtime, earn extra money
  - ***qf12_3_e***: Borrow from family or friends
  - ***qf12_3_f***: Borrow from employer/salary advance
  - ***qf12_3_g***: Pawn something that you own
  - ***qf12_4_k***: Use authorized, arranged overdraft or line of credit
  - ***qf12_4_l***: Use credit card for a cash advance or to pay bills/buy food
  - ***qf12_5_m***: Take out a personal loan from a financial service provider (including bank, credit union or microfinance)
  - ***qf12_5_o***: Take out a loan from an informal provider/moneylender
  - ***qf12_6_p***: Use unauthorized overdraft
  - ***qf12_6_q***: Pay my bills late; miss payments
  - ***qf12_7_r***: Other
  - ***qf12_97***: Don't know
  - ***qf12_99***: No answer
- ***qf13***: If you lost your main source of household income, how long could your household continue to cover living expenses, without borrowing any money? (1 - Less than a week; 2 - At least a week, but not one month; 3 - At least one month, but not three months; 4 - At least three months, but not six months; 5 - More than six months; -97 - Don't know; -99 - No answer)
- ***qk3***: Assume that you receive a gift of €1,000. Imagine that you have to wait for one year to get it and inflation stays at 1%. In one year's time will you be able to buy: (1 - More than you could today; 2 - The same amount; 3 - Less than you could buy today; -97 - Don't know; -99 - No answer)

- ***qk4***: You lend €25 to a friend one evening and he gives you €25 back the next day. How much interest has he paid on this loan? (Numeric answer; -97 - Don't know; -99 - No answer)
- ***qk5***: Suppose you put €100 into a <no fee, tax free>savings account with a guaranteed interest rate of 2% per year. How much would be in the account at the end of the first year, once the interest payment is made? (Numeric answer; -97 - Don't know; -99 - No answer)
- ***qk6***: And how much would be in the account at the end of five years [remembering there are no fees or tax deductions and you don't make any further payments into this account and you don't withdraw any money]? (1 - More than €110; 2 - €110; 3 - Less than €110; 4 - It is impossible to tell from the information given; -97 - Don't know; -99 - No answer)
- ***qk7_1 - qk7_3***: I would like to know whether you think the following statements are true or false: (1 - True; 0 - False; -97 - Don't know; -99 - No answer)
  - ***qk7_1***: An investment with a high return is likely to be high risk.
  - ***qk7_2***: High inflation means that the cost of living is increasing rapidly
  - ***qk7_3***: It is usually possible to reduce the risk of investing in the stock market by buying a wide range of stocks and shares

10.2 Feature Engineering matrix and rules

	Financial knowledge	Indebtment	Perceived fin. knowledge	Saving attitude	Planning attitude	Financial products experience	Independent fin. approach	Digitalisation
qf1	0	0	0	0	1.0	0	1.0	0
qf2	0	0	0	0	0.5	0	0	0
qf3_1	0	0	0	2.0	0	0.3	0.5	-0.5
qf3_3	0	0	0	2.0	0	0.5	0	0
qf3_4	0	0	0	2.0	0	-0.5	-1.0	-0.5
qf3_6	0	0	0	1.0	0	1.5	0	0
qf3_7	0	0	0	1.0	0	1.0	0	0
qf3_8	0	0	0	-2.0	0	0	0	0
qf4	0	0	0	0.5	1.0	0	1.0	0
qf8	0	0	1.0	0	0	0	0	0
qf9_1_9	0	0	0	0	1.0	0	0	0
qf9_2_3	0	0	0	0	1.0	0	0	0
qf9_4	0	0	0	0	1.0	1.0	0.8	0
qf9_5	0	0	0	0	1.0	0	1.0	0
qf9_6	0	0	0	0	1.0	0	1.0	0
qf9_7_8	0	0	0	0	0	0	-1.0	0
qprod1c_1	0	0	0	0	0	0.6	0	0
qprod1c_2	0	0	0	0	0	0.9	0	0
qprod1c_3	0	1.0	0	0	0	0.4	0	0
qprod1c_5	0	0.5	0	0	0	0	0	0
qprod1c_6	0	0	0	0	0	1.0	0	0
qprod1c_7	0	0	0	1.0	0	0.3	0	0
qprod1c_8	0	0	0	1.0	0	0.6	0	0
qprod1c_10	0	0	0	0	0	0.1	0	0
qprod1c_11	0	0	0	0	0	1.0	0	0
qprod1c_12	0	0	0	0	0	1.0	0	0
qprod1c_14	0	0	0	0	0	0.4	0	0
qprod2	0	0	1.0	0	0	0	0	0
qprod3_1_16	0	0	0	0	0	0	0	-0.3
qprod3_2	0	0	0	0	0	0	0	-0.3
qprod3_3_6	0	0	0	0	0	0	0	1.0
qprod3_4	0	0	0	0	0	0	-0.5	-0.2
qprod3_5_7_12_14	0	0	0	0	0	0	0	-1.0
qprod3_8	0	0	0	0	0	0	-1.0	-0.1
qprod3_9	0	0	0	0	0	0	-1.0	0
qprod3_10	0	0	0	0	0	0	-1.0	0
qprod3_11	0	0	0	0	0	0	-0.5	0
qprod3_13_15	0	0	0	0	0	0	0	-0.1
qprod3_17	0	0	0.5	0	0	0	0.7	0
qf10_1	0	0	0	0.5	0	0	0	0
qf10_2	0	0	0	0	-1.0	0	0	0
qf10_3	0	0	0	-0.5	0	0	0	0
qf10_4	0	-0.25	0	0	0	0	0	0
qf10_5	0	0	0	0.5	1.0	0	0	0
qf10_6	0	0	0.25	0	0.5	0	0	0
qf10_7	0	0	0	0	1.0	0	0	0
qf10_8	0	0	0	-0.5	0	0	0	0
qf10_9	0	0.25	0	0	0	0	0	0
qf10_10	0	0.25	0	0	0	0	0	0
qf10_11	0	2.0	0	0	0	0	0	0
qf10_12	0	0	0.25	0	0	0	0	0
qk1	0	0	1.0	0	0	0	0	0
qf11	0	0	0	-2.0	0	0	0	0
qf12_1_a	0	0	0	0.5	0	0	0	0
qf12_1_b	0	0	0	1.0	0	0	1.0	0
qf12_1c_3g	0	0	0	0	0	0	1.0	0
qf12_2d	0	0	0	0	0	0	1.0	0
qf12_3e	0	1.0	0	0	0	0	-1.0	0
qf12_3f	0	1.0	0	0	0	0	-1.0	0
qf12_4k_5m_5o_6p	0	1.0	0	0	0	0	-1.0	0
qf12_4l	0	0	0	0	0	0	1.0	0
qf12_6q	0	0.5	0	0	0	0	0	0
qf13	0	0	0	0.25	0.25	0	0	0
qk3	1/3	0	0	0	0	0	0	0
qk4	1/3	0	0	0	0	0	0	0
qk5	1/3	0	0	0	0	0	0	0
qk6	1/3	0	0	0	0	0	0	0
qk7_1	1/3	0	0	0	0	0	0	0
qk7_2	1/3	0	0	0	0	0	0	0
qk7_3	1/3	0	0	0	0	0	0	0

How to read and use the matrix: let A be a respondent and $\hat{S_A}$ its transformed score for all questions (rows) of the matrix above, M. Then, the vector of values of all new features V (present as matrix columns) can be computed as $V = \hat{S_A} * M$. Note that $\hat{S_A}$ is NOT exactly the vector of score S_A for all question but has instead to follow certain rules for its transformation:

- score for question *"qf8"* is changed as following $\hat{S_A}(qf8) = \bar{X} - X_i$ where $\bar{X}$ is the average of possible scores and X_i is the score of the i-th row; the same applies to scores from *"qf10_i"* and *"qprod3_i"* $\forall i$, *"qk1"* and *"qf1"*
- for *"qf13"*, *"qf2"*, *"qf1"*, *"qf4"*, *"qf11"*, *"qk1"*, *"qf8"*, all *"qf12_i"*, all scores become zero when they are -99 or -97 or 6 (in the case of *8*)
- for *"qf4"* if the score is 2 it becomes zero; moreover, if the score is used to compute values for the independent financial approach variable then if the score is zero it becomes -1
- for *"qf11"* if the score is zero it becomes 0.5
- for all *"qf9_i"* features when included in the planning attitude variable the scores are summed and when positive becomes one, otherwise stays zero. The same applies for the features *"qf12_1_b"*, *"qf12_1c_3g"*, *"qf12_2_d"* and *"qf12_4_l"* when included in the independent financial approach variable; and in the same financial approach variable also this applies to features *"qf12_3_f"*, *"qf12_3_e"* and *"qf12_4k_5m_5o_6p"*

So, to obtain the transformed $\hat{S_A}$, one applies all rules to the simple S_A score.

References

1. Lella, F. "Conti correnti svuotati con un clic, allarme in Puglia: è la quinta regione in Italia. E in tanti si ritrovano indebitati senza saperlo". *Corriere del Mezzogiorno.* https://bari.corriere.it/notizie/cronaca/24_settembre_01/conti-correnti-svuotati-con-un-clic-allarme-in-puglia-e-la-quinta-regione-in-italia-in-tanti-si-ritrovano-indebitati-66eb1381-4012-41ef-81d3-ec118796fxlk.shtml?appunica=true&app_v2=true (2024/09/01).
2. CONSOB. "Truffe e abusivismi - Educazione finanziaria". *Autorità italiana per la vigilanza dei mercati finanziari.* https://www.consob.it/web/investor-education/truffe.
3. Di Salvatore, A., Franceschi, F., Neri, A. & Zanichelli, F. Banca d'Italia - Eurosistema. https://www.bancaditalia.it/pubblicazioni/qef/2018-0435/index.html?com.dotmarketing.htmlpage.language=1 (2018/06/15).
4. Chicco, D. "Ten quick tips for machine learning in computational biology". https://link.springer.com/article/10.1186/s13040-017-0155-3 (2017).
5. Chicco, D. & Jurman, G. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". https://link.springer.com/article/10.1186/s12864-019-6413-7 (2020).
6. Kahneman, D. "Thinking, fast and slow". https://archive.org/details/thinkingfastslow0000kahn_b1q8/mode/2up (2011).