NBC Challenge:

The problem is a binary classification.

In the first step, I tired to do some exploratory data analyses (EDAs):

- Finding irrelevant values - in a number of dummy variables (age, gender_male, number of children, previous loan)
- Replacing null values
- Doing outlier detection - Try to avoid overfitting by removing noises
- Converting categorical variables to dummy variables
- Checking correlation matrix to find the most correlated variables.
  e.g. mono_product and TEF can be dropped to avoid multicollinearity problem while using logistic regression.
  e.g. withdraw dummy variables for bank products are highly correlated with the product volumes (as expected).

The next step, after randomly splitting the sample into test and train (%30 and %70) and do feature scaling, a simple logistic regression has been run. AUROC was about 0.61 and the data model is poor in this case.
I used Logistic L1 with different penalties to do some feature reductions, because it can help to have less complicated model and improve the performance. If the model has large number of parameters then it's going to have high variance and low bias. so, this helps to find a better balance without overfitting and undercutting the data.
But in this case the model did not get better for different scenarios.

Then, I used a random forest. Because it is a collection of decision trees and is much more robust than a single decision tree.
Running an RF with all the variables slightly improved the model. So, I tried the model with dropping variables have less importance.

After several runs, the best AUROC is 0.64, and the importance of variables in the way that two classes to be distinguishable enough is as follows:

	variable	importance
10	revenu_foyer	0.301714
5	vol_transaction	0.114333
1	relation_BNC	0.108741
8	age	0.105085
9	revenu	0.099555
3	vol_epargne	0.062280
4	vol_financement	0.061622
2	vol_MC	0.059764
7	nb_enfant	0.036497
0	nb_produit	0.031395
6	logement	0.019014

The reason I chose AUROC for evaluating the model is that the accuracy is not a good measure of assessment in the case of an imbalanced sample (need to interpret precision and recall).

Further steps to take:

- Initial insight is that the class 1 is only about %10 of the sample. This dataset is an imbalanced sample. Oversampling and undersampling (most of the time) lead to a better data model, the most common one is SMOTE. But the fact is that the there are some disadvantages to these methods.
- The neural network and MLPs might lead to a better data model.