1- The Dataset

The originally received dataset contains 4 Excel sheets: "Soc_Dem", "Products_Actbalance", "Inflow_Outflow" and "Sales_Revenues". Therefore, we decided to use the client identifiers in order to merge these sheets into a single dataset focusing on customers with available sales and revenue information. This data set yielded 969 customers as well as 36 columns.

2- Exploratory analysis and "Dirty Modeling"

a- Exploratory analysis

After having set up our dataset, we decided to explore it to see what it could provide in terms of information.

One of the first things we observed is multiple missing values in different columns. Moreover, by plotting a boxplot, we quickly realize that there are outliers and we have to find a way to reduce their impact on the dataset.

b- "Dirty Modeling"

Unfortunately, we don't have a reference to which we can compare the precision of our models. Consequently, we decided to test 8 different models on our dataset without establishing any specific pre-processing or tuning and comparing the performance indicators in order to choose a **top 5** of models that could help us to solve our challenge.

Model	Accuracy	Precision	Recall	F1-score	AUC*
Logistic Regression	0.81	0.84	0.94	0.89	**0.6489**
Random Forest	0.81	0.81	0.98	0.89	**0.5853**
Nearest Neighbour	0.75	0.79	0.93	0.85	0.5114
Support Vector Machine	0.79	0.79	1.00	0.88	0.5119
Stochastic Gradient Descent	0.74	0.80	0.90	0.85	0.5339
Naïve Bayes	0.74	0.81	0.88	0.84	**0.5684**
Decision Trees	0.68	0.81	0.77	0.79	**0.5634**
Neural Network Model	0.78	0.78	0.99	0.88	0.4867

***Table 1** : Dirty Models and Performance*

As a result, 5 models stood out for their accuracy, precision and the AUC score which is very often used to measure the accuracy of quantitative tests :

- **Logistic Regression**
- **Random Forest**
- **Naïve Bayes**

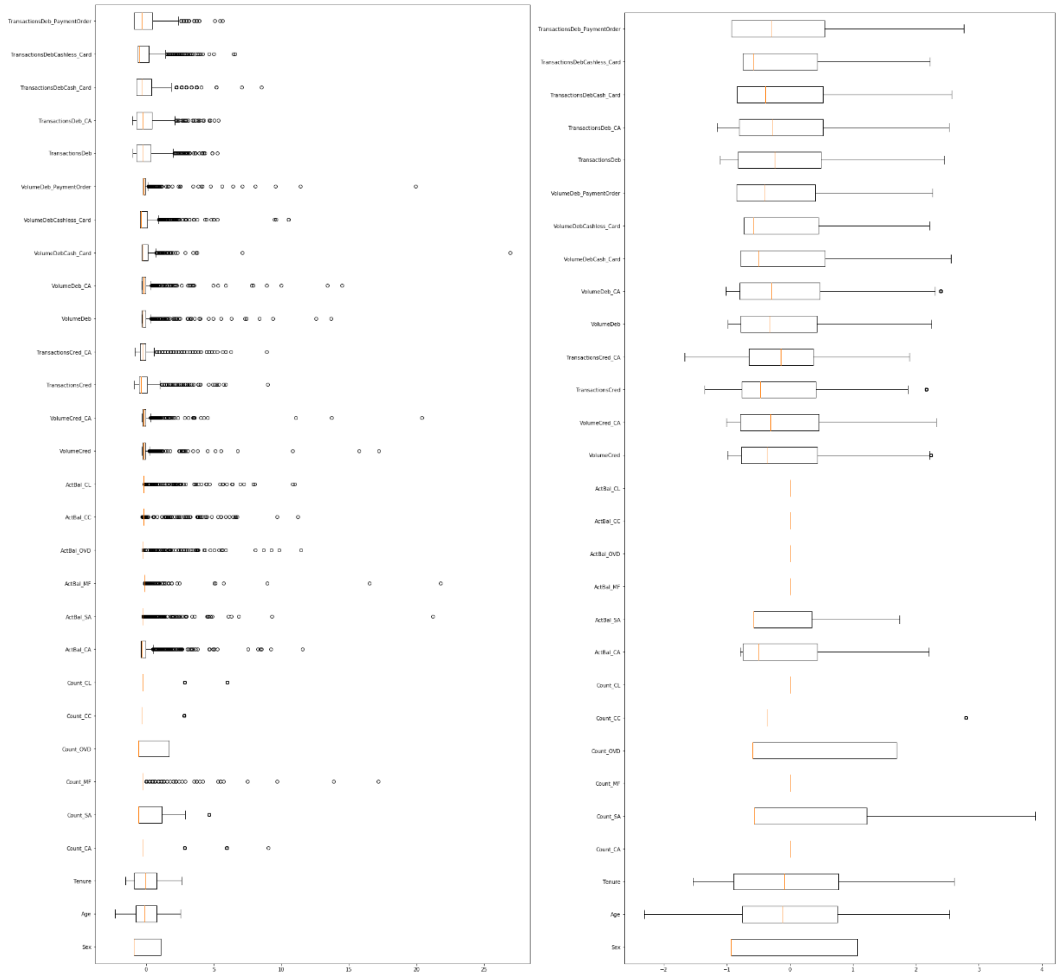
- **Decision Trees**
- **Neural Network Model**

3- Pre-processing

First, we had to handle the missing observations. We saw that there exist two missing observations in "Sex" and multiple missing values in other columns. We tried different alternatives : we substituted the missing values by the mean, mode or median. Thus, we saw that replacing them by 0 and substituting the missing values in the sex column by the mode was the most beneficial option for our study.

In addition, as mentioned above, we had to deal with the problem concerning the outliers. After some research we discovered the truncation function that allowed us to solve this issue.

Before **After**



Representation 1 : *Outlier problem solved*

4- Building the models

We structured our data, solved the issues that we had and defined 5 models that could help to find answers to our questions and this for each product offered by the bank. We have selected, based on

our intuition, different features that could help us to answer the issue. We adjusted these features during the whole process in order to find the ones that would generate more value and accuracy.

Therefore, the models built for each product are :

- **Random Forest for Customer Loan.**

	precision	recall	f1-score	support
0	0.83	0.93	0.88	152
1	0.57	0.31	0.40	42
accuracy			0.80	194
macro avg	0.70	0.62	0.64	194
weighted avg	0.77	0.80	0.78	194

Representation 2 : *Classification report of the model*

- **Neural Network Model for Credit Card.**

	precision	recall	f1-score	support
0	0.71	0.91	0.80	134
1	0.45	0.17	0.24	60
accuracy			0.68	194
macro avg	0.58	0.54	0.52	194
weighted avg	0.63	0.68	0.63	194

Representation 3 : *Classification report of the model*

- **Decision Trees for Mutual Fund.**

	precision	recall	f1-score	support
0	0.85	0.78	0.81	157
1	0.30	0.41	0.34	37
accuracy			0.71	194
macro avg	0.57	0.59	0.58	194
weighted avg	0.74	0.71	0.72	194

Representation 4 : *Classification report of the model*

5- Target Clients

Here are the results of all our models. We decided to select clients who had more than **60%** chance to generate income through the subscription to a particular product. We can therefore target **78** individuals and this could generate **635,6855** in income.

	Consumer loan	Credit card	Mutual fund	**Total**
Number of clients	11	19	50	**78 individuals**
Sum of revenue	65.255714	180.75214	389.677679	**635.6855**

Table 2 *: Target Clients and Revenue*