# ML project formulation

November 20, 2023

## 1 Definitions

- $P_i$ is the probability of default of customer $i$, this is a parameter derived from previous models
- $A_i$ is the amount asked by customer $i$, this is a parameter
- $X_i$ is equal to 1 if the loan is approved, this is a variable
- $T$ is between 0 and 1 and is the threshold to approve or not the loan based on the probability of default, this is a variable
- $C_i$ is the probability that client $i$ accepts the loan considering the interest rate proposed, this is a variable
- $I_i$ is the interest rate proposed by the bank if the loan is approved, this is a variable
- $Z_i$ is the probability that client $i$ accepts the loan considering the interest rate proposed if the loan is approved and 0 otherwise, this is a variable greater than 1

## 2 First formulation

$$
\begin{aligned}
\max_{X,T,C,I,Z} \quad & \sum_i Z_i * (A_i * I_i * (1 - P_i) - A_i * P_i) \\
\text{s.t.} \quad & T - P_i \leq X_i \ \forall i \\
& P_i - T \leq 1 - X_i \ \forall i \\
& Z_i \leq A_i \ \forall i \\
& Z_i \leq C_i \ \forall i \\
& Z_i \geq C_i - (1 - A_i) * M \ \forall i \\
& Z_i \geq 0 \ \forall i \\
& C_i = (I_i - 0.01) * \frac{1}{1} \ \forall i \\
& 0.01 \leq I_i \leq 1 \ \forall i
\end{aligned}
$$

# 3 Second formulation

Once we will have trained our model, we will know the threshold to apply to decide whether the loan is approved or not.

We will then have 2 tests to run :

- Use the merged dataset part that has not been used before to predict the probability of default, decide if the loan should be approved or not and compare to previous applications to check if we have the right threshold
- Use the approved loan from the previous dataset and the default probabilities to determine the interest rate and compare to the one that was actually given

Formulation for second test :

$$\begin{aligned} \max_{C,I} \quad & \sum_i C_i * (A_i * I_i * (1 - P_i) - A_i * P_i) \\ \text{s.t.} \quad & C_i = (I_i - 0.01) * \frac{1}{1} \ \forall i \\ & 0.01 \leq I_i \leq 1 \ \forall i \end{aligned}$$

# 4 Datasets

We will have to split the dataset in 4:

- A training part to train the prediction default models
- A validation part to choose which model/combination of models to use to predict the default
- A second training part on which we predict the probability of default and we use to train the optimization model
- A test part to run our 2 tests

# 5 Notes

It would be logical that the probability of default depends among other things on the loan amount of the application. However, this amount will not be the same in the previous applications, so if we predict for client A a probability p of default, and then apply our tests on their previous application to see if we should approve the loan or not, as the loan for the previous application is different, this probability does not apply anymore: we cannot say that there is a probability of default per person, it is per person and per loan.

Two things we could do :

- Predict the probability of default and then run a feature importance analysis to assess how the amount of the loan impacts the probability of default
- Instead of defining a probability of default p, we could define a probability of default p(amount)