Exercise #1

In the attached "Ex1 - Modeling sample.csv" you will find a dataset containing the following variables:

- **Loanid**: Loan identification number
- **Model1**: Model #1 output – Probability of default
- **Model2**: Model #2 output – Probability of default
- **Target:** Models' target – Binary variable: 1- Defaulted loan; 0- Current loan
- **Sample:** Name of the sample to which each observation belongs (training/ testing/ validation)

Considering that:

- A model needs to be activated in order to accept/ reject clients when applying for a loan within Portfolio A
- Model #1 and Model #2 have been built using data from Portfolio A
- Portfolio A's acceptance rate after the model is activated should be 80%
- Once the model is activated, Portfolio A's owner will evaluate its performance using only loans repayment information collected AFTER the model was activated

What is the best model to be activated for Portfolio A? Why?

Exercise #2

In the attached "Ex2 - Data sample.csv" you will find a dataset containing the following variables:

- **ID**: Loan identification number
- **Fold**: Name of the sample to which each observation belongs (IS = training/ OS = testing)
- **Target:** Models' target – Binary variable: 1- Defaulted loan; 0- Current loan
- **Var4/ Var5/ …/Var163**: Potential predictors

Additionally, we are sharing a dictionary where you will find the type of each of the potential predictors (Continuous/ Categorical Ordinal/ Categorical Nominal/ Dummy).

Considering that, we ask you to:

- Train models to predict the outcome *Target* given the values of the predictors using the IS (training sample) only.
- Test your models' performance against the OS (testing sample) and choose the model you consider is the best performing one to predict *Target* given the available data.

Please consider that:

- Candidate models do not necessarily need to be built using all potential predictors. However, if feature selection is performed, we ask you to explain the taken approach.
- You should not assume that that features have any particular distributions, but you are free to assume any parametric distribution if that makes modeling easier for you.

- We do not expect you to do feature engineering using the potential predictors as variables explanation or economic sense is not being shared. However, predictors can be transformed or combined in any way you think is best to perform this task.
- You are free to modify the training sample in any way you think is best for the process (oversampling, subsampling, splitting on multiple different populations, etc), but you should never use observations in the testing sample to train the model.
- We are open to answer any questions you have while you are working on this exercise, so please do not hesitate to contact us via email if necessary.

We ask you to share with us:

- Any code used for model building and testing, as well as for any additional analysis you perform to solve this exercise
- A document or simple presentation explaining the modeling approach and showing the results and metrics considered to choose the final model
- A dataset including the predictions for the final model

We will rate your performance on this exercise by:

1. Clarity of your code (you may use any tools at your disposal) and the description of the process
2. Thought process and approaches taken
3. Model complexity and overall performance (given the usual performance we expect to obtain from this type of data)